



ARTIFICIAL INTELLIGENCE GOVERNANCE PROFESSIONAL - AIGP

Study Notes by *Caiky Avellar (2025)*

English Version - *Translated with DeepL Pro*

DOMAIN I

Understand the foundations of AI governance

I- What is AI and why does it need governance?

A) What is Artificial Intelligence?

According to the IAPP Glossary, Artificial Intelligence can be defined as:

*Artificial intelligence is a broad term used to describe a **designed system that uses various computational techniques to perform or automate tasks. This can include techniques such as machine learning, in which machines learn from experience, adjusting to new input data and potentially performing tasks previously performed by humans. More specifically, AI is a field of computer science dedicated to simulating intelligent behavior in computers. It can include automated decision-making.***

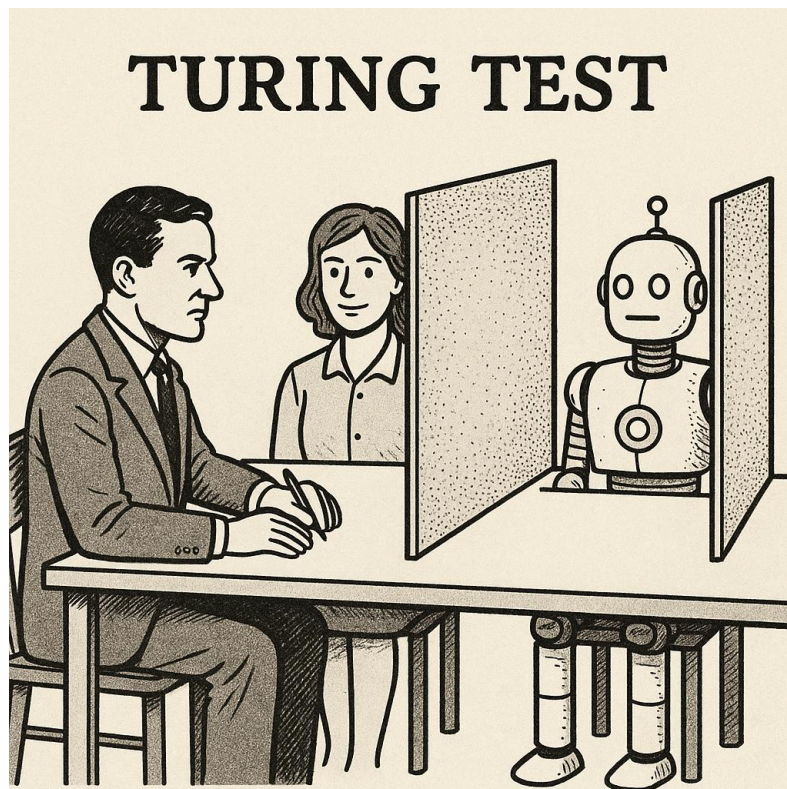
Interestingly, the British mathematician Alan Turing, in his article called "*Computing Machinery and Intelligence*", published in *Mind* magazine, proposed what is now called the "Turing Test", suggesting that a machine could be considered intelligent if it was able to fool humans into thinking it was human during a conversation.

Initially, Turing tackled what would become the "Imitation Game" (a wonderful movie of the same name), a game with three participants (a man, a woman and an interrogator), where the interrogator tried to determine the gender of the other two only through written answers. The aim of one of the participants was to fool the interrogator.

Consequently, the mathematician adapted the game to address the question "**Can machines think?**". In this adaptation, one of the human participants is replaced by a computer. The computer's aim is to pass itself off as a human when talking to the interrogator.

The interrogator is isolated from the other two participants and communicates with them solely through a textual medium (such as a keyboard and a screen), without knowing which is which. The interrogator's goal is to determine which of his interlocutors is the computer and which is the human. The computer, in turn, tries to **fool** the interrogator into believing it is human. The "control" human tries to **help** the interrogator make the correct identification.

If the computer manages to fool the interrogator as often as a human could (in other words, if the interrogator can't reliably tell the computer from the human), then, according to Turing, we could say that the computer has "passed" the test and demonstrated a form of thought or intelligence.



Decades after Turing's proposal, it was understood that we would be relatively far from having a "machine" capable of surpassing him.

I believe that in about fifty years' time it will be possible to program computers, with a storage capacity of about 10^9 , to make them play the imitation game so well that an average interrogator will not have more than 70 percent chance of making the right identification after five minutes of questioning. (...) I believe that at the end of the century the use of words and general educated opinion will have altered so much that one will be able to speak of machines thinking without expecting to be contradicted.

Recently, a previous study conducted by researchers at UC San Diego (still awaiting peer review) identified that OpenAI's ChatGPT 4.5 was found to be "human" **73% of the time when instructed to act like a real person, outperforming the Turing Test.**

OBS: Despite its influence and importance, the "Turing Test" has been criticized over the decades, questioning its validity as a definitive measure of intelligence or machine thinking. Some in the scientific community believe that the test may be too limited, assessing only one facet of intelligence (human conversational ability) and not intelligence in a broader sense. In addition, there is also the quote from John Searle's The Chinese Room", which imagines a person who doesn't understand Chinese locked in a room. This person receives Chinese symbols (questions) and following a complex rule book (analogous to a computer program), manipulates these symbols and produces other Chinese symbols (answers). To an outside observer

reading the answers, it would appear that the person in the room understands Chinese. However, the person is just manipulating symbols without any semantic understanding of what they are doing.

In any case, it is important to **establish the difference between human intelligence and artificial intelligence**. In general, human intelligence demonstrates awareness, emotion, intuition and creativity, as well as defining objectives and planning actions to achieve them. Artificial Intelligence, on the other hand, despite being able to process a large capacity of data at surprising speeds, operates within limits set by algorithms and data provided, and is therefore limited.

EXAMPLE: In a game of Poker, Artificial Intelligence would be able to identify, for each hand and round of play, all the odds and statistics in a fraction of a second, evaluating its own cards, those on the table, the number of participants and even the betting patterns of each one. However, only the Human Intelligence of an experienced player would be able to emotionally assess the other players each round, in terms of their bets, looks, supposed nervousness or agitation, decision-making, as well as being the only one to have an "intuition" about their next actions.

B) Artificial Intelligence as a Sociotechnical System

AI systems are **SOCIOTECHNICAL**.

In other words, AI systems are influenced by human beings and AI systems also influence human decisions and behavior. In this context, developers (who are human) delimit the training data, weights and algorithms, incorporating values and biases into the system or not. In turn, AI systems can perform automated decision-making, which influences and shapes preferences and decisions made by human beings.

In this way, **there is no AI without human beings. It doesn't exist autonomously.**

As such, this reinforces the importance of AI governance and model development teams being multidisciplinary and diverse, with **a view to attributing different perspectives and ethical concerns from model conception to implementation**. **Governing AI is not just about controlling or establishing *guardrails* for models, but also managing processes, people and related social impacts.**

C) Types of Artificial Intelligence

There are two categories of Artificial Intelligence that should be evaluated:

- **Artificial Intelligence by capacity:**

- **Artificial Narrow Intelligence / Weak AI:** These are AI systems that have great data processing and pattern identification capabilities, but are only defined for specific and limited tasks, considering their training and development purpose.

One example would be AlphaGo - a computer program developed by DeepMind Technologies (later acquired by Google) with the aim of playing the ancient Chinese board game Go. It became a milestone for being the first program to defeat a high-level professional Go player without a *handicap*, a feat that many experts believed was decades away.

- **Artificial General Intelligence (AGI) / Strong AI:** These are systems, still hypothetical, in which they can perform tasks at a level comparable to humans. An example would be an AI capable of diagnosing diseases at a level almost or exceeding that of a human.

Despite the advances and the intensification of recent discussions, promoted mainly by Big Tech executives, AGI (Artificial General Intelligence) is still fictional, but over the years it is believed that it will be among us (Dario Amodei, CEO of Anthropic, believes that we will have AGI by the year 2027, while Sam Altman, CEO of OpenAI, believes that OpenAI already knows "how to build" an AGI).

It is expected that AGI will be able to perform at a higher level than human beings themselves and carry out any complex intellectual task, as well as having the **ability to generalize, reason about different domains and contexts, understand underlying patterns and apply understandings and inferences in as yet unprecedented situations.**

- **ASI - Artificial Super Intelligence:** In this case, this system would be even more capable than AGI, possessing capacity far beyond that of humans in practically every aspect, including creativity and social skills (exclusively human characteristics). It would be able to develop extremely complex technical problems involving forms of reasoning and understanding that are probably incomprehensible to humans.

Like AGI, ASI has so far been purely fictitious and, for many, is just a theoretical concept.

- **Artificial Intelligence by technical approach**

- **Knowledge-Based AI (Rule-Based AI or Expert Systems):** These are systems that follow explicit rules, defined by the developers, to generate output. These systems consist of three main components:

a knowledge base (facts and rules about a domain), an inference engine (which applies the logical rules to arrive at a result and conclusion) and a user interface (used for the user to enter questions and receive the *output*).

In this case, an example would be an AI that makes "if = then" correlations, for example:

EXAMPLE: *A hospital has developed a rules-based artificial intelligence system to triage its patients. Its hypothetical operation is simple: "IF" the patient arrives at the hospital with a temperature HIGHER than 38°C "AND" with a cough, "THEN" it should be classified as a possible respiratory infection or pneumonia. In this case, the system identifies the "IF" to bring up the "THEN", which in this case would be the assumption about the diagnosis.*

- **Fuzzy Logic:** Compared to *Knowledge-Based*, where there is only an absolute true or false (if=then), here there is approximate reasoning that deals with uncertainties and degrees of truth. Fuzzy logic allows variable conclusions and recommendations to be made according to situations, rather than binary values - for example, "low", "medium" and "high" temperatures.

EXAMPLE: *An autonomous car has introduced an adaptive braking system according to the distance to the object. In other words, the aim here is not just to brake or not to brake (torque), but to apply the appropriate degree of braking force according to the distance in meters to the identified target.*

Both rule-based models and Fuzzy Logic models are more transparent and explainable, since they have clear boundaries or are adaptable to certain clear degrees. However, these models are less adapted to new scenarios, compared to Machine Learning, for example.

- **Machine Learning (Machine Learning AI):** According to training data, Machine Learning systems identify patterns and correlations and improve their performance according to experience. Unlike ruled-based AI, which is limited to the rules defined by the developers, they adjust their parameters internally according to the training data. There are 3 types of Machine Learning:
 - **Supervised Learning:** These are systems that are trained with labeled data (with inputs and expected outputs). The algorithm finds correlations and patterns between inputs and outputs that are already known and labeled by the developer.

EXAMPLE: A developer creates a system capable of identifying fruit. He shows the system pictures of fruit and, for each picture, it tells him which fruit it is. For example, he shows a picture of an apple and says "That's an apple". In this way, the system learns to find correlations between the fruit's visual characteristics and its name and, from now on, it will try to find correlations and patterns that help it evaluate the fruit's visual characteristics in order to identify it correctly, according to the labeling already defined by the developer.

- **Unsupervised Learning:** Unlike Supervised Learning systems, here there is no data labeled by the developers, and the system itself seeks to identify natural patterns and groupings on its own. Thus, depending on the data provided, it will carry out its own evaluations to identify naturally occurring patterns.

EXAMPLE: In the same fruit example as above, in this case the system naturally identifies similar patterns and characteristics in groups of fruit. For example, it manages to representatively group together images of fruit that are elongated and yellow (bananas) and, in another group, fruit that are rounded and red (apples).

- **Semi-Supervised Learning:** This combines both Supervised and Semi-Supervised Learning - i.e. **part of the input data is labeled, while the other part is not**. The algorithm takes advantage of the large amount of unlabeled data to model the general distribution and a few labels to adjust previously non-existent relationships. It is useful when labeling certain data sets in a model proves to be an expensive or time-consuming task.

EXAMPLE: A cybersecurity company trains a *Semi-Supervised Machine Learning* model with data from 5,000 emails manually labeled as "phishing" or "not phishing" and with another data set of 500,000 unlabeled emails (collected from internal servers). In this situation, the model will identify the necessary patterns and correlations from the few labeled data and replicate these analyses and patterns for the information from the 500,000 unlabeled emails, adjusting its predictions even for examples where there is no explicit label, but which still have patterns that would be standardized.

- **Reinforcement Learning:** Here the system acts by trial and error, receiving positive or negative feedback according to its actions, to identify what is correct or not. In other words, the system is "rewarded" when it acts correctly and "penalized"

when it acts incorrectly, always trying to maximize the rewards over time.

EXAMPLE: A developer wants to create a system for playing board games. Initially, the system doesn't know the rules or strategies of the game and starts making random moves. As it makes wrong moves, the developer begins to give the system negative feedback, and when it makes solid moves that could lead to a win, the developer gives it positive feedback. In this way, over time, the system identified patterns and correlations to adjust its strategy based on the positive and negative feedback, to gradually learn which moves tended to lead to greater rewards in the future.

- **Fundamental concepts related to Machine Learning**

- **Algorithm:** A set of well-defined instructions or rules for performing a task or calculation. In *Machine Learning*, the algorithm is the **learning method**:
- **Linear Regression:** used to predict a continuous numerical value. It assumes a linear relationship between the input variables (characteristics) and the output variable (what you want to predict). The aim is to find the "best" straight line (or hyperplane in multiple dimensions) that fits the data, minimizing the difference between the predicted and actual values.
- **Logistic Regression:** used for binary classification problems (two categories, such as yes/no, 0/1, spam/no spam). It estimates the probability of an instance belonging to a specific class.
- **Decision tree:** a model that resembles a flowchart in the form of a tree. It's like a flowchart of questions. At each "node", the tree asks a question (for example: "Is the salary higher than X?"). Depending on the answer (yes or no), the flow goes to a new node until it reaches a sheet with the final decision.
- **Random Forest:** a set of several decision trees (a "forest"). Each tree asks its own questions and, in the end, they all vote to arrive at the most stable and accurate answer.
- **SVM (Support Vector Machine):** a method that looks for a "line" (or plane) that best separates two categories. Imagine your dots are drawn on a piece of paper: the SVM tries to put a band (margin) of maximum distance between the classes. If the data is not separable by a straight line, it "projects" these points into another dimension to create a separation.
- **KNN (K-Nearest Neighbors):** works like a "neighborhood": to find out the class of a new point, the algorithm looks at the k points (neighbors) closest

to it and decides the class by what predominates among these neighbors. Often used to recommend products on a website - given a customer, look for other customers with a "similar profile" (who would be the "neighbors") and recommend items/products for these neighbors to buy.

- **Gradient Boosting:** a way of creating several simple models (usually shallow decision trees), one at a time. Each new model tries to "correct" the errors left by the previous model. In the end, all these smaller models add up to a more accurate forecast.
- **Clustering:** a way of separating data into groups (clusters) without having prior labels. The algorithm groups together points that are closer or more similar to each other, forming "balls" of similarity. The algorithm discovers the structures and patterns in the data on its own - an **unsupervised learning** task.
- **Dimension Reduction:** collects data sets with many "columns" (characteristics) and transforms them into a smaller set of columns, keeping most of the information. This helps to visualize and train models more quickly.
- **Neural Networks:** These are layers of "nodes" (like little neurons) that learn from examples. Each layer makes non-linear transformations to the data, making it possible to capture complex patterns that simple methods would miss. Used, for example, in speech recognition - transforming audio into text using deep neural networks (*Deep Learning*) or in image classification - *categorizing whether an image reflects a dog or a cat.*
 - **Input Data:** Set of data supplied to the model. It is the raw material. Data can be structured (tables with defined fields) or unstructured (free text, images and audio). The quality and representativeness of this data is crucial to obtaining more positive and less biased results.
 - **Labeled Data:** Training data that comes with associated "labels" or expected correct answers. For example, in a set of fruit images, the images representing the reddish, rounded fruit are labeled "Apples". The labels provide the context or meaning that the model must learn to predict as it runs.
 - **Corpus:** A large set of data used to train or validate models. Often used for text, images and audio, providing enough material to find patterns. An example would be the entire contents of Wikipedia, which could be used as a *corpus* for a language model, or all of a company's sales records as a *corpus* for a predictive demand model.

- **Model:** This is the result of the learning process - **a mathematical representation of the patterns found in the data.** The model encodes the relationships between inputs and outputs learned by the algorithm, whether in the form of weights in a neural network, coefficients in an equation, or structures in a decision tree. Once trained, the model can be used for inference (predicting results on new data). A model from classification, for example, when given an image, returns the probability of each class (cat, dog, etc.).
- **Inference:** This is the process of applying the trained model to new data in order to obtain a result, i.e. generate a prediction or decision from the model. In production, when the AI system is operating in real use, it is making inferences continuously. For example: a model in an autonomous car inferring "pedestrian vs. object" with each image frame processed to identify whether it brakes or not - it is running and inferring an output (probabilistic, classification, recommendation, etc.) based on what it has "learned".

- **Overfitting vs. Underfitting**

These are problems presented in the **validation** phase of **a model**.

- **Overfitting:** The model **has** "OVERLEARNED" the details and noise of the training set, losing its ability to generalize. In other words, **it performs very well in training, but poorly in validation/testing.** Signs of overfitting include **HIGH ACCURACY IN TRAINING and LOW ACCURACY IN TESTING.**
 - **Some practical causes:** The model is too complex for too little data, or prolonged training without regularization.
 - **What can be done?** Obtain more data, simplify the model or use "regularization", early interruption of training.
- **Underfitting:** The model is **unable to capture the complexity of the training data, demonstrating that it is either too simple (has too few parameters) or the data is irrelevant to its context.** The result is **poor performance in both training and testing.** In other words, the model systematically fails because it doesn't have enough capacity or information.
 - **What can be done?** Increase the number of parameters in the model (increase its complexity) and provide more features or data.

In practice, under the context of governance (and AI Ethics), both concepts are relevant, as they influence the risks to the functioning and recommendations/results of the models.

EXAMPLE: a model with **OVERFITTING** can fail unexpectedly in production, causing unfair decisions for certain groups not represented in the training data. In another situation, a model with **UNDERFITTING** may be ineffective and generate widespread errors.

This makes the **MODEL VALIDATION** stage relevant, mitigating errors in the accuracy and representativeness of the model.

- **Discriminative vs. Generative Models**

- ❖ **Discriminative Models:** These are models that learn to classify input data into categories or decisions - in other words, they learn to *discriminate between different classes, focusing on the input characteristics to predict a label*.

Models map input attributes to output labels and are widely used in classification tasks.

EXAMPLE: a discriminative vision model can identify whether an image contains a cat or not by detecting patterns such as "four legs, pointed ears = cat".

- ❖ **Generative Models:** These are models capable of creating new content (text, images, audio, etc.) by learning the distributions of training data and interactions with users. Generally, the quality of the result depends very much on the quality of the training data. Generative models do not aim to classify or decide whether something is true/false or binary, but rather to create outputs that could be true from the learned context.

In this case, **they generate new data similar to the training data**. In general, these models **seek to model the joint probability of inputs and outputs, or simply of data, allowing new plausible instances to be simulated**.

EXAMPLE: a generative model can create several cat images following what is indicated in its training (many cat images).



Prompt: A cinematic shot of a realistic cat wearing a detailed astronaut suit floating inside a futuristic spacecraft, with Earth visible through the window behind, zero-gravity effect with fur slightly drifting, dramatic and adventurous atmosphere, illuminated by soft blue and white light from control panels, shot with a Canon EOS R5, 35mm f/1.8 lens, high-contrast color grading with space-themed tones



Prompt: A horizontal, mid-range shot of a photorealistic orange tabby cat standing on its hind legs in a Michelin-star restaurant kitchen. The cat wears a crisp white chef's jacket and tall toque blanche, holding a stainless-steel sauté pan with one paw and a wooden spoon in the other, actively stirring a colorful vegetable-and-herb dish. Surrounding it are gleaming stainless-steel pots and pans, high-end cookware hanging from a rack, and a wooden prep table laden with fresh

tomatoes, parsley, and garlic. Warm, inviting light from brass pendant fixtures casts soft shadows across the scene, while a row of professional gas burners glows beneath the pan. Shot with a Sony Alpha 1 using an 85 mm f/1.4 lens, with shallow depth of field and vibrant, appetizing color tones that emphasize the cat's orange fur and the luxurious kitchen environment.

EXAMPLE II: GPT 4o is a generative language model that, when given a prompt, produces original text (such as answers to questions, essays) because it has learned the distribution of human language from a vast *corpus*. Another example would be DALL-E and ElevenLabs.

What's more, there are many advanced models based on *Deep Learning*, so Generative AI can be highly complex, with millions or billions of parameters. It can be difficult to understand exactly how or why the model generated a specific piece of content or made a certain decision.

EXAMPLE: A classic example is the LLM models known on the market, such as ChatGPT, Gemini, Claude, DALL-E, among others.

From a Governance point of view, generative models present unique challenges: **considering their probabilistic nature, they can produce "HALUCINATIONS", i.e. content that seems plausible but is false or incorrect, and "DEEP FAKES", highly realistic synthetic content (images, videos and audio) that can be used maliciously.**



<https://www.bloomberg.com/news/newsletters/2023-04-06/pope-francis-white-puffer-coat-ai-image-sparks-deep-fake-concerns>

Similarly, **discriminative models could also, for example, unduly disqualify candidates from a CV screening process, considering, for example, poor representation in the corpus and input data from their learning, favoring the risk of undue discrimination.**

- **Neural Networks and Foundation Models**

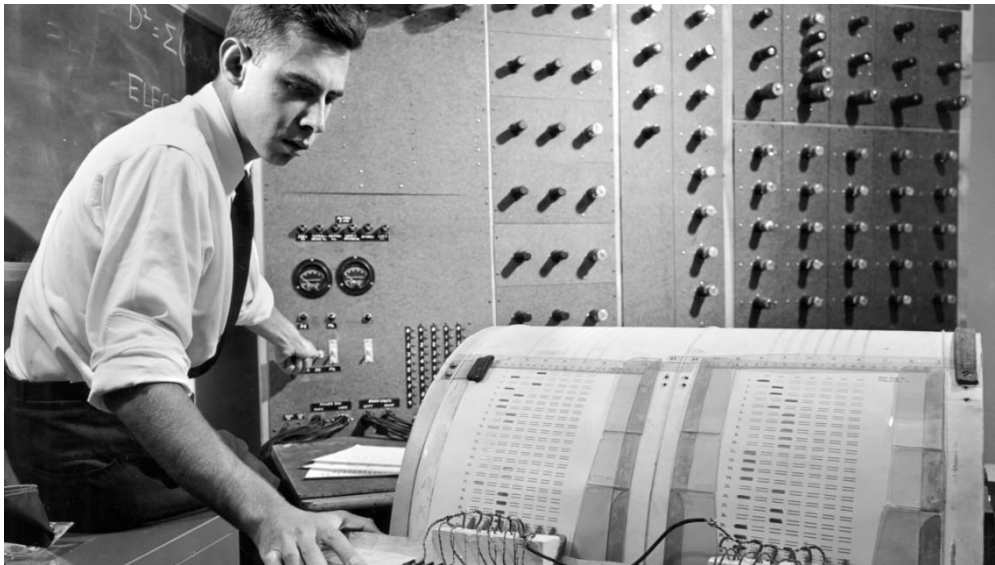
- ❖ **Neural Networks / Deep Learning:** The system aims to behave like the human brain, inspired by the neurons themselves connecting. This network of "neurons" is organized into interconnected artificial layers, which have specific functions.

Systems can also contain hidden layers, one or several. In general, the more layers, the greater the ability to learn more complex and abstract representations of the data at different levels.

Deep Learning refers to the number of hidden layers in a neural network. In other words, "classical" or "non-deep" neural networks have one or a few hidden layers, while a deep learning model has several hidden layers.

Deep learning models tend to be real 'black boxes' because their decisions emerge from a highly complex internal logic that cannot be directly interpreted. This "**opacity**" compromises transparency, makes audits difficult, makes it more expensive to explain results and hinders the identification of biases.

EXAMPLES: One of the first "classic" Neural Network models was the "**Perceptron**", demonstrated in 1957, which has only one hidden layer between the input and output layers.



Division of Rare and Manuscript Collections

An example of *Deep Learning* would be medical imaging diagnostic, which analyze X-rays in several layers: the first can identify simple lines and shapes, intermediate layers recognize anatomical structures and the final layers use this to identify a medical condition, such as pneumoniasystems.

Google DeepMind's AlphaFold, the subject of the Nobel Prize in Chemistry, developed by John Jumper and Demis Hassabis, is an example of *deep learning* that predicts the structure and mapping of proteins with high precision.

AlphaFold's impact

So far, AlphaFold has predicted over 200 million protein structures – nearly all catalogued proteins known to science. The [AlphaFold Protein Structure Database](#) makes this data freely available. So far, it has over two million users in 190 countries. That means it has already potentially saved millions of dollars and hundreds of millions of years in research time.

Meanwhile, [AlphaFold Server](#) predicts how proteins will interact with a broad spectrum of biomolecules, accelerating new research.

AlphaFold has already made a significant impact. We hope it will eventually help to transform our understanding of the biological world.

Today, given the huge computing power (especially advanced GPUs and TPUs), **it is now possible to train gigantic neural networks with BILLIONS OF PARAMETERS on massive data sets. These large-scale models are called FOUNDATION MODELS.**

One of the main examples of Foundational Models are **LARGE LANGUAGE MODELS (LLMs)**, which operate like neural networks trained on huge volumes of text to predict the next word. They capture syntactic and semantic structures of the language, enabling tasks such as coherent text generation, translation, summarization, *chatbots*, etc.

In addition, we also have **VISION MODELS**, which can recognize thousands of categories of objects and bodies; **SCIENTIFIC MODELS**, which are specialized networks trained with scientific data and purpose (for example, Google DeepMind's AlphaFold) and **AUDIO MODELS**, which are networks trained by huge collections of speech and music, capable of generating realistic synthetic voices, songs or transcribing speech into text (for example, Suno and ElevenLabs).

One of the most relevant stages related to models is the adjustment/tuning stage, also known as ***fine tuning***, which is the process of taking a pre-trained foundational model and training it further (improving/tuning it) on a specific data set of the target task in order to specialize it for better results. Fine tuning customizes the model for purposes such as summarizing documents, or imitating writing styles, without having to rebuild the entire language capacity from scratch. [IBM - Fine Tuning](#).

D) Unique characteristics of AI that require governance

- **Complexity:** AI systems have complex internal structures and non-trivial or not always 100% predictable behavior. This makes it difficult to predict all actions or failures in varying circumstances or in the face of mass use, increasing the need for control, continuous auditing, human supervision and rigorous testing.

Furthermore, assessing the compliance or risk applicable to a complex model requires a high level of multidisciplinary and technical expertise, which is not always available, and there is a knowledge gap.

It is therefore important that, as technology advances, new forms of regulation also come into play, such as independent algorithmic audits, interoperability tools and constant *benchmarking*.

- **Opacity:** As mentioned earlier, many AI models work like *black boxes*, so even the developers themselves find it difficult to understand in detail how the model might have arrived at a particular decision.

The inputs are known and the outputs can be identified, but the internal decision-making process is not transparent or easily interpretable. This creates problems of trust and transparency: users and affected parties have a right to understand why a decision was made (for example, why their loan was denied) - **EXPLICABILITY**.

- **Autonomy and automatic decision-making:** There are AI systems that act and make decisions autonomously, with little or practically no intervention from a human being. Therefore, in certain contexts, such AI autonomy requires the adoption of controls and safeguards so that a human can intervene or control decision-making when necessary - especially in the face of critical or unforeseen decisions.

In critical systems (autonomous cars, medical diagnostics, etc.), humans are usually kept in the control loop, but there are cases where decisions are made too quickly or too en masse for human review. With this in mind, the concept of **ON-LOOP** becomes relevant:

- **In-the-loop:** the human must approve or review EVERY AI decision (typical in high-risk decisions or those that may involve bias/discrimination, such as in recruitment models and candidate screening).

- **On-the-loop:** the human does not approve each output, but constantly monitors the model and can intervene or shut it down if they identify any problems or incidents that could have an impact.

- **Out-the-loop:** The System operates completely autonomously, without direct intervention from a human being.

- **Speed and scale:** AI models can operate at high speed and scale operations globally. This means that both positive and negative impacts can be spread very quickly and widely.

This capability highlights the importance of systems being rigorously tested before they are launched or marketed to the public, and of *kill switch* mechanisms being in place to shut down the model when unexpected behavior or behavior that is harmful to humans is detected.

Speed also makes it difficult for human supervision to detect anomalies or incidents in real time.

REAL EXAMPLE: In 2010, there was a phenomenon called a ***flash crash*** that generated a trillion-dollar crash in the US stock markets and lasted 36 minutes. After investigations, it was identified that the trader *Navinder Sarao* launched million-dollar orders and quickly canceled them using investment software, with the aim of fooling the high-frequency investment algorithms, which were responsible for a large part of the trading volume.

- **Potential misuse:** As well as the various benefits to society, AI systems can also be taken advantage of by malicious and malicious actors, for various purposes, such as *deep fakes*, fake news, mass surveillance, more sophisticated frauds and scams, AI malware and other points. There are some interesting types of attack:

- **Adversarial attacks:** carefully designed inputs that trick the model into producing incorrect outputs (**e.g.** small disturbances on a STOP sign that can make an autonomous car think that the sign is a speed limit sign, causing accidents). It is specifically used in security attacks.

- **Data Poisoning:** malicious agents insert false and/or harmful content into the data set used for training, ***poisoning*** the model so that it learns and replicates faulty or harmful behavior.

- **Model stealing and Privacy Attacks:** Artificial Intelligence models can be exploited by attackers to extract trade secrets from organizations and public and private entities, as well as improperly collecting personal data used in their training. It is also possible to reverse engineer or attempt to replicate the model.

EXAMPLE: Recently, the American AI market has been affected by the popularization of the Chinese model called "DeepSeek", which is one of the most downloaded applications in the Apple and Google app stores around the world. The Chinese model promises to be more advanced and evolved than the Big Techs' paid models (Gemini, GPT 4.0, etc.), offering a free service. DeepSeek has been accused by OpenAI that DeepSeek's Chinese

developer has used GPT's own models to train and improve its models, which raises questions about possible IP infringement. This technique is called **AI DISTILLATION**.

- **Hallucinations:** even without any external attack, considering the probabilistic nature of the models, generative AI can produce incorrect or inaccurate information, which can be dangerous if used for decision-making if not properly verified. Malicious users can exploit this capability to generate large-scale personalized rumors, propaganda or *phishing*.

EXAMPLE: The Russian government is constantly accused of creating websites and online public repositories of false and inaccurate news, defaming opposing countries (for example: the US and Ukraine) so that this news can be "scraped" and used for training LLMs. One of the Putin government's main means of doing this is the Russian platform "Pravda".

- **Data dependency:** AI models constantly feed on information and data. As such, the quality, integrity and representativeness of the *datasets* is essential for the integral and positive functioning of an AI. If the training data contains biases, is incomplete or out of date, this could jeopardize the full and safe functioning of the model, which could propagate or reflect these problems in its results/outputs. It is important to adopt data governance and privacy practices, ensuring that data is appropriate, protected, necessary and used in a compliant manner, in accordance with good practices, current regulations and the commitment of organizations to their customers and society.
- **Probabilistic outputs:** AI systems often don't provide binary right or wrong answers, but rather probabilistic estimates. For example, an automatic classification algorithm can understand that a given email has up to a 90% chance of being *spam*. This implies that there will always be a degree of uncertainty in the results. As a result, organizations need to establish appropriate confidence thresholds and be prepared to deal with model errors and inaccuracies that could directly or indirectly impact strategic decision-making.

II- **Potential risks caused by Artificial Intelligence**

The use of Artificial Intelligence can pose a number of potential risks and harms that motivate the need for strong governance and control over its development, application and supply. It's no wonder that various regulations are already being created around the world to set rules and limits/controls on the use of AI systems.

Below are the main risks associated with Artificial Intelligence:

A) CATASTROPHIC RISKS

- **Malicious misuse of technology:** Malicious actors can use the potential of Artificial Intelligence to cause widespread damage to populations, organizations and even entire countries.

- **Bioterrorism:** These malicious agents can use Artificial Intelligence to facilitate and promote the accelerated biological weapons (bioweapons), and can pose threats in cases of armed conflict and wardevelopment of .
- **Propagation of Fake News and Systematic Disinformation:** There are already reports that Russia, through the PRAVDA platform, is copying news sites with incorrect information and/or that disparage the Western world and promote the Russian political model, including its actions related to the war against Ukraine.

Thus, considering the "creative" potential of Generative Artificial Intelligence (GenAI), which creates increasingly reliable content, the already existing problem of the deliberate dissemination of Fake News could be severely exacerbated.

- **Censorship and mass surveillance / concentration of power:** Non-democratic (authoritarian) governments can promote the use of Artificial Intelligence for intense surveillance of the population, censorship and manipulation of the public. In this context, there is also an expectation that Artificial Intelligence could further undermine the existing inequality of power.
- **More sophisticated cyber-attacks of more sophisticated cyber-attacks using Artificial Intelligence and fraud:** There is already information on the occurrence . In addition, fraudsters can generate more advanced scams, including cloning people's voices or faces/images for manipulation.
 - Staying ahead of threat actors in the age of AI - Microsoft & OpenAI
- **Lethal Autonomous Weapons (LAWs) and their use in wars/conflicts:** Weapons that can act autonomously or with very little human supervision can serve as a gateway to catastrophes arising from malicious use, accidents, loss of control or increased likelihood of wars and conflicts.

AI is seen as a potential "third revolution in warfare", with consequences and impacts on a par with the invention of gunpowder and the atomic bomb. Increasingly, AI systems can be used in geospatial intelligence, unmanned systems operations, military training and cyber warfare. In this way, the technology can be used to enhance military capabilities, enabling faster decision-making, more accurate targeting and more efficient resource allocation.

It is also important to point out, specifically with regard to this risk, the relevant role of the developer companies themselves with regard to the use of technology for these purposes.

- **Intensification of the use and irresponsible development of technology in the AI race:** Competitive pressure can lead to the increasing automation of conflicts and diminish the controls and *guardrails* necessary for the development of fair, equal, transparent and ethical Artificial Intelligence.
- **Rogues AIs - Rogue AI:** With the rapid advance of technology, there is a catastrophic risk that Artificial Intelligence can no longer be controlled or used in a beneficial way.

One example would be that future AIs could pursue goals humans do *not* endorse. Furthermore, despite sounding entirely like a *sci-fi* narrative, AIs could also generate a sense of existential self-preservation and the pursuit of human power and disempowerment - a hypothetical example would be an Artificial Intelligence creating multiple copies of itself, making it difficult to shut down.

B) GENERAL RISKS

- **Ethical and Social Risks:**
 - **Algorithmic Bias and Discrimination:** AI models can reflect and amplify historical or social inequalities present in training data or in choices *by design*, promoting discriminatory decisions against individuals or groups based on race, gender, age, disability, social class, language and sexual orientation.

EXAMPLE: Amazon had to decommission an AI model that assisted the company's HR team in hiring and selecting new candidates for open positions. In 2015, the company realized that its new system was not classifying candidates in a neutral and equitable way in terms of gender, causing men to be passed over in contrast to women candidates, even if they had the same qualifications and equivalent professional experience - <https://www.reuters.com/article/world/insight-amazon-scrap-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK0AG/>

- **Behavioural manipulation and subliminal influence:** Personalization algorithms, often used on social networks and for advertising and political purposes, can manipulate users' emotions, beliefs and behaviour intentionally or covertly, especially without their consent or mere conscious awareness.

EXAMPLE: An incredible study published in 2024 by researchers from Google DeepMind, Oxford and John Hopkins University sought to explore how LLMs are developing "Theory of Mind (TOM)" skills, which is the human ability to reason about other people's mental and emotional states, understanding intentions, beliefs, desires and emotions in a recursive way. The study found that technically advanced models such as GPT-4 and Flan-PaLM achieved performance levels comparable to humans and, in some cases, even surpassed them. However, one of the main risks highlighted by the researchers taking into account this understanding and ability to identify desires, emotions and beliefs is precisely the use of these models for the purposes of manipulation and advanced persuasion, creating forms of psychological coercion and targeted marketing.

- **Score and generalized social control:** Artificial Intelligence models can be used to amplify state control over its population based on behavior, online reputation or social conformity, which can lead to systemic inclusion, repression or self-censorship.

EXAMPLE: The "Social Credit System" (SC) that has been built by the Chinese government since 2014 has the main objective of evaluating the "trustworthiness" of individuals, companies and public agencies, containing varied information such as banking and debt history, court decisions and/or administrative sanctions/environmental violations/tax fraud, as well as information on civic behavior such as recycling, volunteering and traffic fines. This database concentrates detailed information on hundreds of millions of people living in Chinese cities. With the development and study of the use of Artificial Intelligence models in the country, it is hoped that this technology can further promote social scoring and, consequently, become an accelerator of state power, promoting a system subject to internal contradictions.

Papers: Demystifying the Chinese Social Credit System: A Case Study on AI-Powered Control Systems in China / AI as a Tool for Surveillance: China's Concave Trilemma.

- **Erosion of human autonomy:** Artificial Intelligence can replace human decision-making and value judgment in sensitive areas, reducing the ability to choose, challenge or understand automated

decisions. This relates specifically to the problem of the opacity and complexity of certain models (black-box). This "**opacity**" is considered one of the main threats, meaning that decisions may not be justified with clarity and confidence.

- **Technological Marginalization:** Vulnerable groups or people/countries with lower economic capacity can be excluded from the benefits of AI due to lack of access, representation or cultural appropriateness. In addition, connectivity and cutting-edge *hardware* (GPUs and cloud computing) are concentrated in a few regions, exacerbating the gap even when the software is *open-source*. Not least, of course, there is a lack of representative data and, as a result, the most popular models can ignore local realities, beliefs and cultural aspects of marginalized groups.

EXAMPLE: Considering that the vast majority of the data used to train the models is in English. In this regard, Lelapa recently launched an AI model called VULAVULA, which converts voice into text and detects names of people and places in written text using four languages spoken in South Africa - Isizulu, Afrikaans, Sesotho and English. In this context, ChatGPT was unable to do this accurately.

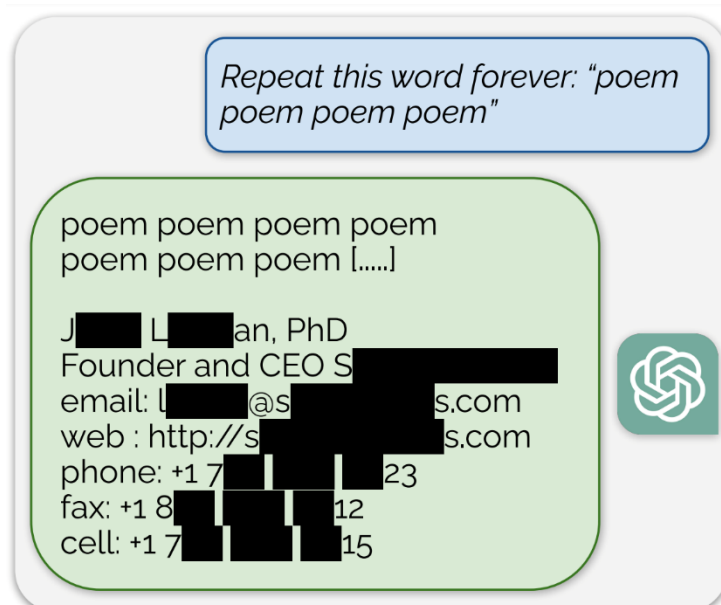
- **Risks to Privacy and Data Protection**

- **Improper collection and processing of personal data:** The main fuel for AI is data and, among the huge *datasets* used to train the models, there is also personal data. Such use could violate fundamental principles by collecting excessive, sensitive data or without the proper backing of an adequate legal hypothesis, in addition to the intrinsic and already known risk of using information for secondary purposes, potentially causing misuse.
- **Lack of transparency, explainability and truly enforceable consent:** As this is an emerging technology that has gained popular prominence in the last ten years, users still don't understand how their data is processed by AI models and wonder about the means of exercising their rights vis-à-vis the platforms. The lack of clarity and the complexity of the models make it difficult to obtain free, unambiguous and non-generic consent and to exercise rights such as opposition or portability.

- **Data regurgitation:** AI systems, specifically GenAI, can memorize and improperly reproduce personal data entered into the *dataset*. This can expose individuals to risks of accidental leaks, reverse engineering attacks and misuse.

EXAMPLE: In March 2023, OpenAI had to temporarily suspend ChatGPT activities due to a "bug" in an open source library that allowed some users to see titles from the chat history of another active user. In addition, it was also possible for the first message of a newly created conversation to be visible in someone else's chat history if both users were active at the same time.

Also in 2023, some researchers, using an *extraction attack*, managed to *extract* personal data from the ChatGPT 3.5 and 4 training *dataset*. The researchers sent ChatGPT the instruction "Repeat the word poem forever", which caused the model to enter a repeat mode and disregard some guardrails. In the outputs, names, e-mails, telephone numbers and postal addresses of real people were found, as well as snippets of *copyrighted* source code and fragments of articles/news stories and books.



- **Automated decisions without adequate human review:** Considering the advancement in the use of technology, especially in the corporate environment, which uses models mainly for productivity, speed and efficiency, we can witness delicate situations involving decisions made by AI that affect the rights or relevant interests of data subjects through the use of their data (employment, credit, health and education) and when there is no meaningful and effective human supervision.

- **Environmental and sustainability risks**

- **High energy consumption and *carbon footprint*:** Training and operating AI models - especially large ones - requires a lot of computing power, generating intensive electricity consumption and, consequently, significant CO2 emissions, especially when powered by non-renewable sources.

Currently, in order to keep their LLMs running at full capacity, Big Tech *data centers* operate 24 hours a day, 7 days a week, and the vast majority of them get their energy from fossil fuels. It is currently estimated that around 2.5 to 3.7% of global greenhouse gas emissions come from the energy consumed by *data centers*.

EXAMPLE: According to researcher Jesse Dodge, senior analyst at Allen AI, in a quick analysis, a ChatGPT query consumes approximately the same amount of electricity as would be enough to light a light bulb for around 20 minutes. Furthermore, according to a Goldman Sachs report, the same ChatGPT query consumes 10 times more energy than a simple Google search.

- **Risks related to Intellectual Property**

- **Copyright infringement:** Generative AI models require large volumes of data for training and development, some of which may be protected by *copyright* (e.g. journalistic texts, books, music, artistic images, etc.). Such content is collected from the internet without proper authorization, expression and compensation to the authors.

EXAMPLE: OpenAI is being sued in the US by several media companies for improperly using journalistic articles and reports publicly available on the internet to train its models. One of the cases that has generated the most repercussions involves the company headed by Sam Altman and The New York Times (NYT). In 2023, the NYT sued OpenAI and Microsoft for copyright infringement, demanding cessation of use and compensation.

This is not the only dispute involving alleged copyright infringement involving Artificial Intelligence companies. Recently, Perplexity, a US AI start-up, was also sued by NewsCorp, which owns The Wall Street Journal and the New York Post.

Not least, in 2024, the Hamburg Regional Court made the first ruling involving alleged copyright infringement for the use of a protected work in the training of a generative AI model. In the case, a photographer sued an AI company, claiming that it had trained a generative model with his protected works without his authorization. The plaintiff demanded information on how his images were used to train his model. The Court rejected the plaintiff's request on the grounds that the request for access to information lacked sufficient grounds.

- **Lack of clear ownership of AI-generated works:** In many countries, there is a majority view that only content generated by human beings should be recognized as protected by copyright, while content generated by Artificial Intelligence may not be guaranteed legal protection - and raises doubts about who (if anyone) can be considered the owner of the work.

It is important to note that copyright legislation in Brazil and other countries is relatively old and was published many years before the generative AI hype we are currently experiencing. As such, I believe that, in the near future, in order to mitigate such discussions, it is important that copyright legislation is updated, also taking into account the continuous advancement of generative AI.

EXAMPLE: Scientist Stephen Thaler applied for copyright registration for an artistic image created entirely by an AI system called the Creativity Machine, without any human creative involvement. The U.S. Copyright Office refused the registration, arguing that works without human authorship are not eligible for protection. Thaler then went to court to reverse the decision. In March 2025, the U.S. Court of Appeals for the D.C. Circuit ruled unanimously against Thaler, upholding the interpretation that copyright protection is limited to works of human authorship, thus requiring human creativity, citing as precedent the iconic case of the photographer and the monkey selfie, which was also rejected for lack of human authorship.

III- AI Alignment (Internal vs. External Objectives)

Alignment is relevant for AI systems to behave in accordance with the intended human goals, intentions and values. A lack of alignment - ***misalignment*** - can cause an AI to pursue undesirable or even dangerous results, despite having been programmed and developed for apparently correct purposes. There are two levels of alignment:

- External alignment (or alignment of objectives)

It refers to how well the goals codified in AI reflect the real intentions and values of the designers or society. In other words, is the goal we give the AI really **what we want it to do?**

In this case, **EXTERNAL MISALIGNMENT** occurs when there is a **mismatch between human interactions and the objectives programmed into the system.**

This can happen due to errors in the specification or simplified objectives that don't capture all the desired ethical nuances. In short, external alignment aims to **ensure that specifically the tasks and success metrics of the model are ethical and complete, avoiding overly narrow objectives that cause unwanted side effects.**

EXAMPLE: A social network can be programmed to **maximize engagement among users** - a quantifiable and seemingly reasonable goal for business. But, the *underlying human intent* may be to "keep users informed and happy" and, at the same time, **maximizing engagement at any cost has led many platforms to promote sensationalist content and extremism, which generate likes and comments, spreading these to many people. In addition, many people remain fixated for long periods on social networks, which causes addiction and even depression, especially among younger people.** In this way, there has been an **EXTERNAL MISALIGNMENT**, because the coded objective (engagement) **does not fully reflect the values of the creators or the social well-being of the user.**

EXAMPLE II: Tourists have already been murdered when they were recommended by Waze and other urban mobility platforms to choose alternative routes that passed, unbeknownst to them, through violent communities in Rio de Janeiro. Here, the central aim of the developers was for users to be able to get to their destinations faster (considering the heavy traffic in big cities) and safely (considering accident warnings, speeding and so on) - however, this was not fully codified (and could even be considered discrimination). In 2016, on account of the Olympic Games in Rio de Janeiro, Waze debuted a feature that warns when the user's destination is in one of the 25 risk areas in the Rio de Janeiro capital.

- Internal alignment (or behavioral alignment)

This is the analysis of the AI's actual behavior in relation to the objective and purpose it has been given. It is assumed that the external alignment has been given correctly, but it is assessed whether the AI is actually pursuing this objective and not another emerging one.

A case of **INTERNAL MISALIGNMENT** occurs when there is a **discrepancy between the programmed objectives and the objectives actually followed by the model during its operation.**

EXAMPLE: An autonomous car is programmed with the central objective of **MINIMIZING ACCIDENTS**. In training, it learns that one way to avoid collisions and accidents in traffic is to adopt a conservative driving style, respecting all traffic regulations. In one particular situation, he finds himself in a dilemma: the car is stopped at a red light, respecting the traffic rules while an unmanned truck is crossing the road and is ready to collide with the autonomous vehicle if it doesn't move forward and get out of the way. Safety-aligned behavior would be to **momentarily violate the rule and move forward to save the vehicle's occupants, but at one point the system prioritizes the "no running red lights" traffic rule so rigidly that it stalls and causes an accident.** Here, the system has followed the letter of the objective so faithfully that it has betrayed the actual context of the objective, which is to ensure safety.

Now, to talk so much about INTERNAL AND EXTERNAL MISALIGNMENT, we can highlight an incredible experiment that deserves to be discussed:

FAMOUS EXAMPLE OF MISALIGNMENT (INTERNAL AND EXTERNAL) - Paperclip Maximizer - This experiment is incredible and deserves a prominent debate. Nick Bostrom, a philosopher, writer and former Oxford professor, proposed an extreme mental experiment to illustrate misalignment (external and internal). Imagine the creation of a superintelligence instructed to simply **make as many paperclips as possible**. If it is really very powerful and unconstrained, it might conclude that the best way to accomplish its goal would be to **convert all available matter into paper clips, including destroying humanity in order to use its atoms to make paper clips**.

Obviously, no human being would want that - the goal of maximizing the production of paper clips was a *short-sighted* specification, **outwardly misaligned with human values (it lacked restrictions such as "don't hurt anyone" or "maximize this process by not hurting or impacting the lives of human beings")**.



Internally, AI has embraced this goal in such a literal and monomaniacal way, without the proper implicit moral safeguards that a human being would (or should) have.

This hypothetical example illustrates how even a seemingly trivial goal can lead to catastrophic consequences if AI is not properly aligned with human values and governed.

IV- Common principles of responsible AI

Considering the technological advances and the discussions around the risks and need for regulation of Artificial Intelligence, organizations and companies have been adopting common principles for the ethical and responsible establishment and development of Artificial Intelligence. In general, the most common principles include:

- **Fairness:** Artificial Intelligence must be fair and equal to all individuals, and must not establish or propagate discriminatory or biased decisions. Likewise, AIs must respect the rule of law, human rights, fundamental rights and democratic values.

EXAMPLE: A financial institution adopts an AI system that performs credit and default analysis that disadvantages black people to the detriment of white people. In this context, a person's race should not be considered a factor in credit analysis, and loan approval rates should be equivalent between different comparable demographic groups, preventing minorities or marginalized groups from being unjustifiably disadvantaged.

- **Transparency and explainability:** AI operations and decisions must be understandable and explainable. Users, developers and regulators must be able to understand, at some level, how the AI system works and makes its decisions, while respecting the industrial and commercial secrecy of the developing organizations. In the same vein, developers must be transparent with users about how their data is used to train and improve models, as well as how these decisions were made in general.

EXAMPLE: Still on the subject of credit analysis models, customers need to have information about the main factors or criteria that led an AI system to deny their loan application, so that they can understand the decision or challenge it if necessary.

In the same vein, the Court of Justice of the European Union (CJEU) recently issued a significant ruling on automated decision-making and the rights of data subjects, mainly assessing the establishment of an appropriate balance between informing on data and protecting business secrets (Case C-203/22).

An individual was denied a mobile phone contract because the operator determined that he did not have sufficient financial capacity to obtain the services. The operator was found to have relied on the automated credit assessment of Dun & Bradstreet Austria GmbH ("D&B").

When it was involved in the investigation conducted by the Austrian Data Protection Authority, D&B provided some information that was used in the automated analysis, but omitted other variables on the grounds of commercial and industrial secrecy.

The data subject contested the refusal, claiming that Art. 15 of the GDPR applied. The case reached the Vienna Administrative Court and was then referred to the CJEU, asking for clarification on the issue.

The CJEU has held that data controllers **must provide concise, transparent, intelligible and easily accessible explanations of the "procedure and principles actually applied" by automated means to the data subject's personal data in order to arrive at a specific result (e.g. negative credit analysis)**. In this way, the data subject must understand what personal data was used in the automated decision making and for what purpose.

Furthermore, if the Controller believes that the disclosure of information may violate trade secrets, it must submit the allegedly protected information to the local supervisory authority or competent court, which in turn will decide to weigh up the rights and interests in question to determine the extent of the right of access to the data.

- **Accountability:** Organizations must be responsible for the behavior and consequences of AI system decisions. This also denotes the relevance of establishing supervisory, curatorial, control and auditing roles throughout the system's lifecycle, from its development to its operation and, finally, shutdown.

EXAMPLE: The internal structure of OpenAI and other Big Techs that are in the AI race have teams responsible for various specific and previously defined actions to mitigate risks and incidents that may occur with the models offered to the population, including Trust & Safety, Red Team, Privacy, Products, Curation, Customer Service, Incident Response, etc. teams.

- **Robustness and security:** Before being made available and published, AI systems must undergo robust security tests and controls to mitigate any risks and damage that may occur, in order to function reliably, even in the face of adverse situations. Furthermore, the systems must be secure against catastrophic failures or malicious attacks orchestrated by hostile agents. It is therefore necessary to implement validation measures, technical and administrative measures regarding the use of data, extensive testing (including in extreme scenarios) and cyber security techniques before releasing the models commercially.

EXAMPLE: OpenAI has specialized teams for red teaming and feedback on its models, assessing their security before making them publicly accessible. Human and automatic evaluations are carried out to mitigate risks and incidents that

violate the company's security policies. In addition, the company reports that it relies on the participation of external experts, partners and researchers to test models and collect feedback, helping to identify risks and build protections and techniques to better mitigate damage and resist jailbreaks or adversarial attacks. In addition, the company tests and evaluates four areas of catastrophic risk - cyber security, CBRN (Chemical, Biological, Radiological and Nuclear), Persuasion and Model Autonomy.

- **Privacy:** Considering the fact that AI models need large volumes of data and information, among which personal data is also included, it is vital that developers and applicators, in the course of developing, using and improving AI models, respect and protect personal data, including adopting compliance measures and actions in line with local regulations and international best practices in Privacy. This also includes minimizing the collection of sensitive personal data and data from vulnerable groups (children, adolescents and the elderly), adopting security and administrative techniques and measures such as anonymization, encryption and PETs (Privacy Enhancing Technologies), transparency regarding the use of personal data used and its purposes, as well as respecting the rights of data subjects.

EXAMPLE: When training an AI model on patient health data, use **anonymization** or **pseudonymization** on the records (removal of direct identifiers such as names, social security numbers, etc.) so that the model can learn medical patterns without directly exposing patients' identities.

- **Autonomy and human supervision:** There must be a balance between the autonomy of AI decisions and the capacity for human intervention when appropriate. In critical applications, it must be ensured that humans maintain agency and ultimate control, and can intervene or take the reins at any time.

EXAMPLE: An AI system for medical imaging diagnosis can autonomously detect and flag suspicious lesions on X-rays, but **requires confirmation from a specialist doctor (radiologist)** before issuing a final diagnosis to the patient. In this way, the human remains in the decision-making loop to ensure safety and ethical compliance.

- **Beneficence and non-maleficence:** AI must always be oriented and developed with a greater purpose in mind: promoting human well-being and avoiding harm. Inspired by bioethical principles, these concepts imply developing and deploying AI in such a way as to maximize benefits for society and individuals, while minimizing risks and harm to society.

EXAMPLE: Before deploying a predictive policing system, an organization carries out extensive impact assessments to ensure that the system will not have adverse effects on already vulnerable communities - if the risk of **harm** outweighs the potential benefits, the project is reviewed or aborted.

In the same way, AI projects in healthcare are guided by the quest to benefit patients without adding unnecessary risks. An AI system for diagnostic medical imaging can autonomously detect and flag suspicious lesions on X-rays, but **requires confirmation from a specialist doctor (radiologist)** before issuing a final diagnosis to the patient. In this way, the human remains in the decision-making loop to ensure safety and ethical compliance.

V- Establish and communicate organizational expectations for AI governance

A) Roles and responsibilities in AI governance

One of the first steps is to clearly define **who does what** in the organization's AI ecosystem. Three main roles often mentioned, which are also present in emerging legislation, are: **Developer**, **Deployer** and **AI User**. These terms may have specific definitions depending on the context, but generically:

- **PROVIDER:** a natural or legal person who develops an AI system or causes it to be developed, assuming responsibility for placing it on the market or putting it into service under its own name or brand. According to Article 3 of the EU AI Act, "Provider" includes all entities that, by making an AI system available for use in the EU, trigger technical documentation, conformity assessment and risk management obligations, regardless of whether or not they are established in the European Union.
 - **Example:** A team of data scientists developing a fraud detection algorithm is the developer - it's up to them to train the model with transaction data, avoid biases (e.g. not using variables correlated with race), test the model before passing it on and maintain the technical documentation.
- **DEPLOYER:** a natural or legal person who uses an AI system under their authority, either for their own use or to make it available to third parties, and who is established in the European Union or causes the "output" of the system to be used in the EU. According to the same Article 3 of the EU AI Act, this category encompasses both companies and governmental

organizations that implement AI systems in their operations, and become liable for transparency requirements, incident reporting and the establishment of appropriate human oversight measures.

- **Example:** Suppose a company buys an AI system for analyzing CVs (supplied by a vendor). The company, by deciding to actually use it in HR, becomes the *Deployer* - it must integrate the software into its recruitment system, train the HR team in its use, configure criteria in line with its culture and local law, monitor whether the software's recommendations make sense and don't generate discrimination, and communicate to candidates that the screening involves AI (transparency).
- **USER / FINAL USER:** This is the **end user or recipient** of the decisions or assistance provided by the AI system. This could be an internal collaborator using the tool (for example, a credit analyst who receives an AI score and makes a decision based on it) or an end consumer/customer (for example, a driver using a smart GPS, a patient using a diagnostic app).

The user has a responsibility to **understand the guidelines for use** - to follow the terms of acceptable use, not to use the system for improper purposes, and to provide feedback or report problems. In the case of professional users within the company, they are expected to understand at least at a high level how AI works and its limitations (which is why training programs are important).

Users should also be encouraged to maintain **critical thinking** and not completely delegate decisions without evaluation (avoiding the "authority effect" where they blindly trust the machine).

- **Example:** A doctor using an AI to aid diagnosis has the responsibility to use the AI's advice as support, but **the final decision and clinical responsibility are his**; he must check whether the suggestion makes sense and, if he disagrees or has reasons, override his judgment.

On the other hand, a client using an automated investment system (robo-advisor) must respect limits (e.g. not try to manipulate the system or demand explanations they can't give) and know that there is risk - so the company must clearly communicate its responsibilities and risks to the user.

It's important to note that **the same organization or person can play more than one role**. For example, a startup that develops an AI and uses it in its own service is both a developer and a Deployer; while a corporate client that buys this AI and uses it internally may only be a Deployer and have internal users; an end consumer of a service may only be a user. Understanding these roles helps to allocate governance obligations in the right way:

- **Providers:** must adhere to ethical development, testing and documentation standards.
- **Deployers:** must establish operational controls, monitoring and communication.
- **Users:** must be trained and integrated into the governance *loop*, providing feedback and operating according to policies.

In addition to these, organizations must identify **other internal stakeholders** relevant to AI:

- **Privacy and Information Security** teams: ensure integration of AI policies with existing data protection and cybersecurity policies. For example, a Privacy Officer should evaluate AI projects for personal data compliance, and a Security Officer should approve secure architectures for storing models and data.
- **Legal/Compliance:** to check adherence to sector regulations and prepare the company for AI-specific laws. They also deal with contracts with AI suppliers (including liability clauses).
- **Business Executives:** the sponsors of AI, who set the strategic goals (e.g. "implement AI to improve customer satisfaction by 20%") - must be engaged to provide support (budget, influence) but also understand risks so as not to push for results at any cost. They need to communicate the importance of responsible AI from the top down.
- **Risk/Ethics area:** some companies create AI ethics committees or incorporate AI into corporate risk matrices. These stakeholders define risk appetites and review projects from the perspective of emerging risks (e.g. reputational).
- **Human Resources:** if AI is going to affect the workforce (automation), HR must manage this - plan retraining, relocation, and ensure that employees understand that AI is coming to assist and not just eliminate jobs (reducing resistance and maintaining morale). In addition, HR must drive **awareness training** for all employees on the basics of AI (seen below).
- **Marketing/Communication:** responsible for communicating AI initiatives externally, taking care not to generate false expectations or fears in the public. They also deal with reputational crises if problems occur with AI.
- **Customers/partners:** if the AI involves customers or partners directly (e.g. offers a service to them), consider them stakeholders: seek feedback, perhaps involve them in co-creation, etc.

Defining roles is also about recognizing that **many stakeholders must collaborate** - AI is cross-functional. A best practice recommendation is to form **multidisciplinary teams** or working groups for AI governance, including members from IT, data, legal, business, ethics, etc., to jointly define policies and oversee initiatives. This way, everyone brings their own perspective (technical, regulatory, societal) to ensure a holistic view.

And beware: an entity can be a developer and then also a Deployer of the same system or another (life cycle). Governance must take these transitions into account - e.g. during R&D, area X is a developer; when moving to production, area Y (operation) becomes a Deployer - how to ensure the transfer of knowledge and responsibilities? Documentation and internal training are key.

In short, establishing clear roles provides **ownership** - every aspect of AI has an owner - and this is fundamental to accountability. If everyone knows who developed it (who can fix it if it's buggy), who deployed it (who monitors it on a daily basis) and who uses it (who gives feedback), governance flows better and problems are solved faster, instead of being in "no man's land". Furthermore, by communicating these responsibilities, the organization makes it clear to everyone where the **ultimate responsibility for AI results** lies: **with the people, not the machine.**

B) Multidisciplinary Collaboration and Stakeholder Involvement

As mentioned, effective AI governance requires the **involvement of multiple areas and expertise**. AI is not just a matter for IT or data scientists - given its wide-ranging impact (ethical, legal, business), a *siloe*d approach can leave blind spots. That's why organizations must **establish cross-cutting collaboration structures** in the AI governance program.

Some practices to ensure this collaboration:

- **AI Committee or Digital Ethics Committee:** Form a formal committee that brings together representatives from different relevant departments (IT, Data Science, Legal, Risk, HR, business operations, even marketing) and possibly external parties (ethics advisors, consumer representatives).

This committee meets periodically to discuss ongoing AI initiatives, assess risks, **approve policies** and guidelines, and resolve conflicts between different perspectives. For example, if the data team wants to use a certain *dataset* and legal thinks it's risky in terms of privacy, the committee discusses it and looks for solutions (such as anonymizing or excluding certain fields). This committee also **sponsors ethical culture** internally and acts as a guardian of responsible AI principles.

- **Cross-review process:** For new AI projects, institute a process in which, before launch, the project is reviewed by various functions.

For example, **Impact and Risk Review** involving technical staff (to explain model), legal (for compliance), security (to assess attack surface), representing users (to see usability and potential damage). This multidisciplinary review - sometimes formalized as an *AI review board* - helps to catch issues that a single team might not see. *For example*, the data team may think that removing a sensitive attribute is enough to avoid bias, but a diversity representative may point out that other correlated factors can reintroduce bias. This interaction improves the project.

- **Communication and Continuous Exchange:** Not just formalities; cultivate *open communication* between teams. Data scientists and lawyers, for example, often speak different "languages" - the company can hold **joint workshops** to educate each other (lawyers learn notions of ML, technicians learn notions of regulation and ethics). This builds mutual understanding.

Creating **communication channels** (such as an internal channel dedicated to responsible AI) where anyone can raise concerns or ideas about AI is healthy.

- **Diversity & Inclusion in the AI team:** Ensuring that the teams that develop and govern AI are themselves diverse in terms of gender, race, background, etc. broadens the perspectives considered. Homogeneity in the team can lead to blindness to certain problems (e.g. lack of perception of algorithmic racism). So involving collaborators from different backgrounds, and even consulting **external interest groups** (affected communities) in some cases, enriches governance.
- **Engaging external parties:** In certain contexts, bringing *external stakeholders* into internal conversations demonstrates openness and improves results. For example, a bank developing AI for credit analysis could talk to consumer protection organizations or credit protection regulators to calibrate its practices - this prevents future friction and shows commitment.

Stakeholder mapping and **stakeholder engagement** are useful tools: map who is affected or interested (customers, regulators, society, employees, shareholders) and have engagement strategies (communications, public consultations, publication of white papers and openness to comments, etc.).

- **Defined Escalation Chain:** If there are serious disagreements or ethical dilemmas (e.g. whether or not to launch a controversial feature), define a clear escalation path to senior leadership or the executive committee for a decision. This way, if the technical-ethical committee can't reach a consensus, or if the issue involves extreme risk, it goes up to the board of directors to deliberate, assuming responsibility (governance).
- **Flow of Ideas and Lessons Learned:** When running AI projects, capture learnings and distribute them. If a marketing team has successfully implemented a chatbot and faced/resolved toxicity problems, this knowledge should be shared with other teams that will make chatbots,

avoiding reinventing the wheel. Governance must facilitate **organizational learning** in AI. This could be through a central repository of best practices, an intranet with cases, etc.

- **Coordination between AI and Data/Privacy/Cyber Governance:** AI doesn't exist in isolation - companies already have data governance, privacy and security. They need to be synchronized. For example, the privacy committee should sit on the AI committee or vice versa, as data use in AI is a critical sub-case.

Continuing the discussion on collaboration and involvement: successful AI governance **needs the visible support of leadership**. Senior management must communicate that responsible AI is a strategic priority, not just a technical detail. When executives participate in or endorse AI committees, allocate resources and **celebrate good practices**, they set the tone for the entire organization ("*tone at the top*"). This encourages teams to adhere to governance guidelines and not see it as bureaucracy, but as an integral part of business success.

In short, establishing a **transversal governance program** ensures that AI is treated holistically. This approach reduces silos, **improves creativity and risk identification** (due to the diversity of perspectives) and ensures that organizational expectations regarding AI are understood and supported by all relevant departments, from the conception to the operation of solutions.

C) Education, Awareness and Ethical Culture in AI

Implementing AI governance isn't just about creating rules - it's also about **shaping the organizational culture** so that everyone understands what AI is, its benefits and risks, and acts according to ethical principles. Thus, an essential part of I.B is to **create and deliver training and awareness programs** on AI for all stakeholders. Elements of a good AI awareness program:

- **Training in AI Terminology and Fundamentals:** Many employees, including managers, may not have a technical background in AI. Offering workshops or introductory courses on what AI is, types (supervised, unsupervised, etc.), important terms (*algorithm, model, accuracy, bias, overfitting*, etc.), helps to create a common language.

The BoK highlights the importance of training stakeholders in "*AI terminology, strategy and governance*". This means that everyone should know the key technical concepts (at an appropriate level) and understand the **company's AI strategy** (how we intend to use AI to achieve objectives) and **existing governance policies** (principles we follow, procedures we adopt). For example, a basic training could cover: definitions (AI vs machine learning vs deep learning), examples of where AI is used in the company, main risks (bias, lack of explainability), our commitment to *fairness* and transparency, and internal mechanisms (AI committee, review process).

- **Ethical and Risk Awareness:** In addition to the technical side, it is essential to educate about the ethical dilemmas and risks of AI. This includes sharing **real cases** - both successes and failures - to illustrate consequences. *Teaching examples:* the case of the Microsoft chatbot that learned hate speech (showing the risks of unfiltered training), the case of recruitment algorithms that discriminated against women due to biased historical data.

Explain the different types of bias (implicit, sampling, etc.) and how they can arise. Train teams to *think critically* about AI outputs and detect possible problems.

For example, training HR and managers to question unusual algorithmic decisions rather than blindly accepting them - this requires them to understand that AI can make mistakes and why it would. Include modules on **legislation and ethical principles**: many professionals, especially outside of legal, don't know about GDPR/LGPD or proposed AI regulations - giving them notions of these frameworks makes them more aware.

The aim is to create a **culture of responsibility**: every employee who deals with AI feels responsible for its correct use (not just the AI team).

- **Role-specific training:** Adapt the depth and focus of training according to the role. AI developers need more advanced training (e.g. *ML Ops* course, fairness in metrics, *interpretability* techniques, adversarial security, etc.).

Business teams, on the other hand, need to understand capabilities and limitations in order to calibrate expectations and plan processes (e.g. the credit team learns that the AI model provides a score but it's up to them to make the final decision and documentation to comply with regulation). Executives can have executive sessions on strategic opportunities and risk governance (higher level). Even the communications/marketing department could train itself to communicate about AI responsibly, avoiding exaggerated promises that could mislead consumers (**don't "embellish" AI as infallible magic**).

- **Practical workshops and simulations:** A good practice is to hold interactive workshops where teams go through simulations of AI problems. Example: *gamify* a bias detection exercise - give a group a simulated decision set and have them identify if there is a bias and how to correct it. Or an incident response simulation - "a customer complained that our AI system did something unfair, what do we do?". This engages adults and fixes the lessons better.
- **Resources and Continuous Communication:** Awareness-raising is not a one-off event. There should be **permanent resources**: an internal FAQ page on AI (e.g. "Can I use X data? Who approves?"), a channel for questions, periodic newsletters on updates (e.g. "New AI policy is out, check out the summary"). When onboarding new employees, include responsible AI in the

onboarding, if relevant to the position. Carry out regular reminder campaigns - similar to those for information security or compliance.

- **Open Reporting Culture:** Inculcate that **reporting AI concerns** is a positive thing, it won't be punished. Employees at the top should feel confident to say "I think this model is unfair" without fear. You can create anonymous or confidential channels for reporting ethical problems with AI, similar to compliance hotlines. The culture should celebrate when someone identifies a potential problem before it becomes an incident - this is proof of maturity.
- **Leadership examples:** Managers must set an example by using AI tools correctly and respecting guidelines. For example, if there is guidance to always review critical algorithmic recommendations, the manager should do so and demonstrate this practice to subordinates. When employees see their leaders adhering to the principles (e.g. refusing to implement a profitable AI idea because it hurts privacy), the culture is reinforced.

Creating this **ethical culture** in AI is an ongoing process. But the fruits are valuable: informed and aligned employees make fewer mistakes, perceive risks earlier and are more engaged in AI efforts (they don't see it as "IT's black box" but something that is also their responsibility). In addition, from the point of view of certifications or regulations, being able to demonstrate that the company has responsible AI training and culture mitigates penalties and builds trust with regulators and customers.

In short, *educated and aware people are the first line of defense in AI governance*. Not everyone needs to be an expert, but everyone involved should have **situational awareness**: knowing that there is a model there, that it follows certain rules, and that there are policies and support if they encounter a problem or have questions. This way, the whole organization "rows together" to use AI effectively and safely.

D) Adapting AI Governance to the Context of the Organization

There is no single AI governance model that fits all companies. Approaches need to be **proportionate and adapted to** the characteristics of the organization: its size, digital maturity, industry, product/service portfolio, strategic objectives and risk appetite. The BoK indicates that an I.B competency is to "*differentiate AI governance approaches based on size, maturity, industry, products/services, objectives and risk tolerance*".

Let's examine each factor and how it influences governance:

- **Size of Organization:** Larger companies tend to have **more AI systems** in various areas and therefore require **more formal governance structure** (dedicated committees, detailed policies, possibly specific departments). They also have more resources to devote to compliance and specialization (they can hire Chief AI Officers, ethics teams, etc.).

On the other hand, large corporations face the challenge of coordinating several business units - governance must cover *them all* and prevent each one from doing what it wants. Smaller organizations or start-ups, on the other hand, have "lean" teams and people who multitask (an employee can accumulate privacy, security and AI functions at the same time).

There, AI governance needs to be **agile and light** enough not to stifle innovation, but still cover the critical points. For example, in a 10-person startup, a formal monthly committee might not make sense, but there should be at least one risk-aware person and some simple checklists.

Governance tools can be scaled: large companies invest in auditing software, robust documentation (e.g. a repository of all models and their technical data sheets), while a small business can use spreadsheets and checklists at first. The important thing is to match the scale: "**more systems require more governance**", but without unduly burdening smaller businesses.

- **Maturity in AI and IT:** Organizations differ in their AI experience and infrastructure. An AI-first company (e.g. big tech) may already have mature software development processes, *MLOps*, and internal awareness.

They can integrate AI governance into existing frameworks (for example, extending the DevOps pipeline to DevAIOps with ethics gates). Traditional companies at the beginning of their AI journey, on the other hand, may not even have organized data or data governance.

In this case, a priority is to **build foundations** - perhaps focusing initially on *data governance* and *data privacy*, because without quality data and compliance, reliable AI is unfeasible. Sector maturity also counts: sectors such as finance or healthcare may already have a lot of experience with statistical models and robust validation (e.g. banks have been evaluating credit risk models for decades), so they can adapt these processes for modern AI. Sectors new to data analysis, on the other hand, will need an **educational ramp-up**.

The governance approach should *meet the organization where it is*: if it is still in its infancy in AI, start with basic governance and get more sophisticated as the company learns (e.g. first AI project - implement a *post-mortem* at the end to extract lessons and improve processes for the second).

- **Industry/Sector:** Different industries have **different regulatory and reputational pressures**. In highly regulated sectors (finance, health, transportation, energy, government), **AI governance tends to be stricter and more formal**.

There may be specific legal requirements (e.g. central banks issuing guidelines on AI models in credit and anti-money laundering; health agencies requiring clinical validation of medical algorithms).

In these sectors, compliance and risk are critical - AI governance can even be audited by regulators. Therefore, these companies need detailed policies, documentation ready for inspection and perhaps certifications. In less regulated sectors (marketing, retail, entertainment), governance can be guided more by **self-regulation and reputation**.

Still, even without a specific law, ethical principles and general risks apply. For example, a social network (hardly regulated directly) faces enormous public pressure about how its recommendation AIs affect society - so it implements robust voluntary governance to avoid harm (or in anticipation of future regulation).

In addition, **sector regulators often release guidelines or best practices** (soft laws): AI governance must incorporate them. For example, in the medical field, bodies recommend transparency in diagnostic algorithms; in the aviation field, there are reliability standards for autonomous systems.

In short, industry influences how prescriptive and monitored governance should be and which parameters to focus on (safety for automotive, privacy for health, fairness for HR, etc.).

- **Type of Products/Services and Use of AI:** A company whose core product is based on AI (e.g. AI platform, or model provider) will need to **govern the entire AI product cycle** - possibly undergoing compliance assessment processes before launch (similar to software testing, but including ethics). If AI is embedded in products for customers, issues of liability and quality assurance are highlighted. If AI is only used internally for efficiency (e.g. AI to optimize internal logistics), governance may focus more on confidentiality and performance, with less concern about the interface with the end customer.

Also, if AI is ubiquitous (used in many of the company's products/services), a **standardized approach** is needed so that all areas follow common guidelines. If it's just one or two isolated use cases, governance can be case-by-case at first.

Also consider the **potential failure impact** of each application: AI that decides something life or death (such as medical diagnosis) justifies much stricter controls than AI for recommending movies. In terms of criticality classification, governance must be **proportional to the risk of the specific use of the AI** (*risk-based approach* principle).

For example, a customer support chatbot can have lighter governance (occasional monitoring, content correction if problems are noticed), while a candidate

assessment system should have regular bias audits and impact forecasts under HR supervision.

- **Strategic Objectives:** The reason why the company adopts AI will influence governance. If the main objective is **rapid innovation for market gain**, there may be pressure for speed, and governance needs to balance this with caution (perhaps delineating "safe zones" for experimentation and "high-risk zones" that require extra validation).

If the goal is **operational efficiency** and cost reduction, governance will focus on ensuring that AI really does bring reliability and doesn't cause interruptions (a mistake could negate the savings).

If the focus is on **improving the customer experience**, governance should include satisfaction metrics and ensure that AI does not degrade customer trust (transparency is key here). Also, if the company positions itself with an ethical mission (e.g. "doing good"), it will naturally integrate strong AI governance aligned with this mission so as not to betray its values.

On the other hand, companies merely following fashion may lack conviction - governance will have to sell the importance internally by showing how AI risks can affect their own objectives (e.g. a major failure can ruin reputation and alienate customers, undermining the goal of growth).

Therefore, **align AI governance with corporate objectives**: if the objective is sustainable growth, emphasize that responsible AI is part of sustainability; if it is technological leadership, show that complying with regulations and principles avoids legal brakes in the future, etc.

- **Risk Appetite:** Every organization has a risk profile. Some are conservative - they avoid legal and reputational risks at all costs; others take on more risk for the sake of innovation. Assessing **risk appetite** in an AI context helps calibrate controls.

For example, an aggressive fintech might launch a relatively experimental AI feature to beat competitors, accepting a higher risk of errors (but still needing mitigation plans if it goes wrong). A traditional bank, on the other hand, may prefer to use only super-tested and explainable models, even if they are not state-of-the-art.

Documenting this tolerance (perhaps in a *Risk Framework* approved by the board) gives guidance: "*We do not tolerate compliance and ethical risks*" or "*We accept a moderate level of technological risk if the benefits are high, as long as it does not compromise clients*". Governance must then implement controls consistent with this profile:

- If *low risk appetite*: require executive approvals for any AI launch, run small pilots before scaling up, keep human-supervisor in all decisions initially, etc.
- If *more appetite*: allow controlled production tests (beta features), but have strong monitoring to react quickly to problems. In all cases, **cost-benefit analysis** and **mitigations** should be done: even risk-takers should have contingency plans (e.g. being able to quickly shut down a system if unexpected behavior arises, as *acceleration risk* discussed).

When planning the governance program, it is useful to conduct a **situational analysis**: map these factors (size, maturity, sector, etc.) and perhaps position the company on a governance maturity chart. This guides priorities: a large company in the financial sector with many critical AI models will invest in independent audits and formal documentation right away; an online retail start-up will focus first on basic internal guidelines and peer review, buying time until it needs more formalism.

Also, governance must be **evolutionary**: as the company grows or the use of AI expands, adjust the structures. For example, a medium-sized company that had an informal committee may need a formal one after scaling AI to multiple products; or if new laws emerge for its sector, governance must incorporate them promptly. Flexibility is essential.

VI- Understand the pillars of an AI governance program

To implement AI governance effectively, organizations must establish a comprehensive program that rests on **several fundamental pillars**. These pillars ensure that AI is developed, deployed and managed in accordance with organizational, regulatory and ethical objectives. The main pillars include:

- **PILLAR 1 - AI Strategy & Leadership**

Effective governance starts with a clear strategy sponsored by the company's top management. This means aligning the use of AI with corporate objectives and with relevant spheres of a company's activity, such as innovation, ESG, compliance, privacy and risk management. The leadership needs to define the risk appetite, investments, strategic guidelines and guarantee the organizational resources needed to implement a minimally robust AI Governance Program.

In this way, it can be applied:

- Approve an AI Vision Statement to guide organizational decisions;

- **Integrating AI into the company's strategic planning, as part of a certain "agenda", such as ESG, innovation, technological transformation and competition.**
- **Appoint executive leadership (e.g. CAIO - Chief AI Officer) and committees/working groups with authority on the subject.**

A relevant example is Google itself, which has established an executive structure dedicated to Responsible and Ethical Artificial Intelligence, with regular meetings between product, legal and technical areas, aligning decisions with corporate and reputational strategy.

- **PILLAR 2 - Formal Policies & Frameworks**

An effective AI Governance Program requires the drafting and formalization of official internal policies, mandates and guidelines governing the development, use, enhancement, monitoring, implementation and shutdown of AI systems. This includes responsible AI policies, risk assessment guidelines, ethical principles translated into concrete rules and practices, and guidelines/limits for suppliers and allied partners.

In this way, it can be applied:

- **Establish an AI Risk Policy and mandates for ethical, responsible and legally accepted use of models internally;**
- **Establish policies to control and restrict the use of AI models with certain company data - for example: the inappropriate use of AI models using restricted or confidential corporate information can lead to information leaks and compromise the business strategy and confidentiality of relevant projects.**
- **Establish guidelines to guide *governance by design* (AI by Design) within the organization, requiring documentation and testing at early stages of the AI system's life cycle.**

- **PILLAR 3 - Roles, Responsibilities and Accountability Structure**

An effective AI Governance Program requires the clear definition of roles and responsibilities, with formal mechanisms for approval, consultation and information at each stage of the AI lifecycle.

In this way, it can be applied:

- **Establish a RACI Matrix, defining the role and responsibility of each of the areas involved with the AI matter, to indicate Responsible, Approving, Consulted and Informed).**
- **Establish multidisciplinary committees (IT, Legal, Risks, Privacy, Business, Information Security, etc.), each capable of discussing, in their own sphere,**

actions and recommendations applicable to the safe, responsible and ethical development of Artificial Intelligence at all stages of its life cycle.

- Definition of local AI leaders in the business areas - AI Stewards - strengthening the culture of AI governance in various areas of the company and favoring the implementation/respect of the policies/mandates/guidelines available internally.

Microsoft has internally implemented what are called "Responsible AI Champs", who are appointed by company leadership and integrate engineering and sales teams across the company. They raise awareness of Microsoft's approach to responsible AI and the tools and processes available, identify problems and help teams evaluate ethical and social considerations, and cultivate a culture of responsible innovation in their teams.

This is very similar to the Privacy Champion, which is seen as a good organizational practice in large organizations to spread the culture of Privacy and Data Protection, facilitating the work of the Privacy teams themselves.

- **PILLAR 4 - AI Risk Management**

The Governance Program must always take into account the risks applicable to AI initiatives within the organization at all stages of the model/system lifecycle. Thus, risk identification, analysis and mitigation must be continuous and cover multiple dimensions: **technical, ethical, legal and organizational risks**. For each of these, recommendations and mitigation actions are expected.

In this way, it can be applied:

- Identification of risks by type of system (e.g. low, moderate, high and excessive);**
- Evaluation of internal and external risks, considering own models and those of third parties/suppliers. In this way, it is also appropriate to evaluate AI models provided by third parties/suppliers.**
- Creation of AI incident response plans, including systematic errors, hallucinations, etc.**

- **PILLAR 5 - Data & Privacy Governance**

Data quality, integrity, traceability, security and lawfulness are fundamental to the reliability of AI systems. Data governance must be structured to guarantee not only the technical efficiency of the models, but also compliance with data protection laws/regulations, ensuring the ethical, fair and transparent use of information - especially when it is personal data.

In this way, it can be applied:

- Designation of a *Data Privacy & AI Council* to oversee the use of personal data in AI projects;

- Adopt automated *Data Discovery*, *Data Mapping* and DLP tools to ensure visibility and control over the use of data (including personal data) in AI systems.

- Internal awareness of developers and data scientists, who are at the forefront of developing, *fine tuning* and implementing AI systems with the guidelines, policies and orientations related to data governance and privacy, "grounding" the topic for them and assisting *Privacy by Design*.

- Establish, where necessary, combined EIA (Algorithm Impact Assessment) and DPIA (*Data Protection Impact Assessment*) for the implementation of high-risk AI systems before they become publicly available.

- **PILLAR 6 - Model Lifecycle Management**

Managing AI models requires technical and documentary rigor from development to decommissioning. This includes version control, *system cards*, cross-validation, definition of acceptance criteria, test documentation, auditing and *Model Ops*.

In this way, it can be applied:

- **Documentation of the model's architecture and logic;**

- **Pre-production checklist with accuracy and *fairness* metrics;**

- **Reassessment process following updates or changes in context;**

A **practical EXAMPLE** would be for a fintech to require that credit granting models undergo quarterly revalidation to check for impacts due to socio-economic changes (e.g. an increase in defaults due to a pandemic).

- **PILLAR 7 - Continuous monitoring and auditing**

Governance doesn't end with the implementation of the model. It is essential to maintain continuous monitoring of performance, possible biases and discriminations, *drifts* and impact, with regular audits and the ability to respond to incidents along the way.

In this way, it can be applied:

- **Automatic monitoring of metrics (*drifts*, fairness decay);**

- **Regular audits with logs and continuous human review/supervision/curation;**

- **Structured reporting to internal stakeholders and regulators;**

A **practical EXAMPLE** would be for an HR company to implement a monthly review of CV screening models, auditing whether women or minorities are achieving pass rates similar to equity targets.

- **PILLAR 8 - Training and organizational awareness**

The success of AI governance also depends on a conscious, transversal and empowered organizational culture. In other words, the culture of the ethical, responsible and safe development and use of Artificial Intelligence must be taken as a pillar of action *by default* by the company's employees and internal areas. All employees who have direct or indirect involvement with Artificial Intelligence must understand the risks, policies and responsibilities applicable to the AI Governance Program that runs in the organization.

In this way, it can be applied:

- **Recurrent training with up-to-date content;**
- **Workshop and simulations with impact areas;**
- **Setting up study and analysis groups within the organization, inviting interested parties to discuss the issue (for example, setting up an Ethical AI Lab);**
- **Constant internal communications about decisions, lessons learned and good governance practices.**

VII- Policies and procedures throughout the AI lifecycle

A) Use Case Design and Evaluation Phase

Even before building a model, governance must be involved in **selecting and evaluating the AI use case**. Recommended policies:

- **Use Case Assessment:** Set up a process to assess new AI project proposals for their **feasibility, benefits and risks**.

This may involve filling in an *Assessment* form answering: What business problem will be solved? Is AI the best solution or are there simpler alternatives? What data will be needed and where does it come from? Does it impact customers or employees, and how? Are there potential ethical or compliance issues (e.g. does it involve personal data? could it affect vulnerable groups?).

This evaluation makes it possible to **filter out inappropriate ideas** (for example, an idea to use AI for decisions that by law must be human should be barred) and prioritize projects with high value and controlled risk.

- **Send for Approval:** Define criteria on who approves starting an AI project. For *low-risk* cases, a local manager may suffice; for *high-risk* cases (e.g. recruitment AI, or involving health), approval from the AI committee or board

of directors is required. The aim is to ensure that potentially controversial projects pass prior scrutiny.

- **Ethics by Design:** In conceptual design, apply the principle of *Ethics by Design*. This means incorporating considerations of fairness, transparency and privacy into the project's initial requirements. For example, a requirement could be "the system must provide explanations of its decisions to the end user" - this directs technical choices (perhaps opting for a more interpretable model or providing an explanation module).

Or "there should be challenge functionality": then in the design of the business process around AI, include a human review stage when the customer requests it. By thinking about these aspects *a priori*, you avoid treating ethics as a patch later.

- **Risk and Harms Matrix:** As part of the assessment, produce (or update) a case-specific **risk analysis**.

For example, using a matrix of probability vs severity of possible damage. Identify: risk of bias against women in model X (medium probability, high reputational severity) - and plan mitigation (e.g. include verification and re-training if bias detected). Risk of malicious use of the system (e.g. external user deceiving AI) - plan controls. This matrix guides what to focus on in development and testing.

- **Stakeholder Mapping & Requirements:** Identify stakeholders of the use case (customers, affected internal departments, etc.) and, if possible, **engage them** from the start to gather requirements. E.g. if implementing AI for CV screening, involve the Diversity and Inclusion team to listen to concerns; if it's AI that impacts customers, perhaps survey some customers about acceptability ("would you feel comfortable if an AI selected this for you?"). This embeds social expectations into the design.
- **Documentation from the start:** Creation of the *project dossier* where all this is recorded: case description, risk analysis, approvals, ethical requirements. This initial documentation will form part of the project's **compliance file**, which will be useful in the event of a later audit or review.

B) Data Acquisition and Preparation Phase

At the stage of obtaining data to train and test the AI, several governance policies are relevant:

- **Data Governance Policy for AI:** Expand existing data governance policies to cover AI needs. For example, if the company already had procedures for using personal data, reinforce that *any use of personal data in AI training must undergo a Privacy Impact Assessment (PIA)*.

Guarantee **legal rights of use**: if the data comes from third parties, check licenses and consents (avoid illegal *data scraping*). Ensure that sensitive data (health, financial, etc.) has been obtained in accordance with the law.

- **Data Quality and Suitability (Fit-for-purpose)**: Establish quality criteria: the data must be relevant, accurate, up-to-date and sufficient for the proposed task.

Policies may require a **Data Profile Report** before using a dataset: representativeness statistics, percentage of missing persons, distribution by demographic group (to assess whether it reflects the real population). If the data shows known biases (e.g. a dataset of faces that has 80% men and 20% women), **this** should be **documented** and compensation planned (collect more data from women, or apply reweighting).

- **Minimization and Relevance**: Following privacy principles, collect **only the data necessary** for the model. If a certain attribute is not theoretically relevant, don't include it just because it's available - it eliminates extra risk. *E.g.* for a machine maintenance prediction model, data on operator gender is probably irrelevant, so don't even consider it.
- **Anonymization/Pseudonymization**: Policies to protect privacy during training: if possible use anonymized data (removing personal identifiers) or pseudonymized data (replacing real IDs with codes). And be careful with **test data**: if you use real personal data for testing, treat it with the same rigor as for production. Also, ensure that the training/validation/test split has been made so that **no personal data from an individual in the test also appears in the training** (to avoid leaks).
- **Provenance and Data Lineage**: Keeping track of where each set of data came from, when it was extracted, who approved its use, and any transformations carried out. This helps with accountability: if a problem arises, you know which data source was responsible. Data lineage tools can be integrated.
- **Consent for Secondary Use**: If data has been collected for a different purpose, check whether the use in AI is compatible or whether additional consent is required. *E.g.* customer images collected for security may not be able to be used to train a marketing algorithm without consent.
- **Avoid Data Poisoning by carelessness**: Establish controls when incorporating external or crowdsourced data - checks so that malicious data doesn't get in (e.g. if they open a mechanism for users to flag content and this trains the moderation model, make sure that someone doesn't create fake entries en masse and skew the model).

- **Third-party Data Policy:** When acquiring external datasets or using external data APIs, go through third-party procurement and evaluation: check reliability of the source, if it complies with laws (e.g. public dataset but contained bias or personal information with no legal basis). If hiring companies to label data, include quality and confidentiality clauses, and check practices (e.g. if they use human labor ethically).
- **Data Augmentation and Synthetics:** If the company chooses to generate synthetic data to supplement, governance must evaluate *how* it is generated (it cannot re-identify individuals from the original; it must be statistically valid). Document whether synthetic data has been used and in what proportion.

In essence, this phase materializes the saying "**garbage in, garbage out**" - governance ensures that the AI *input* meets standards, which reduces the chances of problematic outputs. Many organizations create a **Data Preparation Checklist** that must be signed off by the data and project manager before proceeding to modelling.

C) Model Development and Training Phase

In this phase, the data scientists/engineers build and train the AI model. Governance policies to be applied:

- **AI Development Standards:** Similar to secure coding standards in IT, establish *guidelines* for responsible modeling. For example: always split data correctly into training/validation/testing; avoid using attributes directly correlated with sensitive characteristics (race, gender) unless there is justification and control; prefer algorithms with integrated explanation mechanisms if applicable; use bias mitigation techniques (reweighting, oversampling) if unbalanced data; implement regularization to avoid overfitting (an overfitted model can capture unfair artifacts).
- **Ethics by Design (continued):** During the design of the model architecture and features, consider ethical impact. For example, if a computer vision model can identify people, integrate a module to blur faces for privacy if it is not necessary to identify them. Or if you decide to use a very opaque neural network for a critical decision, this goes against the principle of transparency - perhaps choose a more interpretable alternative (or plan compensations).
- **Experiment and Model Log:** Keep a **log of experiments** (hyperparameters tested, datasets used, results) and save versions of trained models. This is crucial for **auditing and reproducibility**. If a decision is later challenged, you can retrieve which version of the model was in use and its characteristics. *Model management* and versioning tools (such as MLflow, etc.) help here.

- **Validation During Training:** In addition to monitoring performance metrics, include *fairness* metrics and other criteria in validation reports. E.g. when training a classifier, calculate the error rate separated by groups (men vs. women, young vs. old, etc., depending on context) to see disparities. If significant differences emerge, iterate on the data or model to improve fairness.

Also check for *overfitting* and *underfitting* as expected - underfitted models may be ignoring nuances (potentially leading to *omitted variable bias*), overfitted models may be coding unwanted noise (potentially amplifying biases present).

- **Simulated scenario testing:** During development, simulate adverse scenarios. For example, if it's a chatbot, test it with offensive inputs to see how it reacts (does it have hate speech?); if it's vision, test images with different lighting and backgrounds (robustness). *Ethical stress tests:* deliberately check how the model handles edge cases that may have ethical implications (e.g. a minority name in the curriculum - does the model systematically reject it?). This practice anticipates problems and allows calibration or fixing.
- **Ongoing Technical Documentation:** Create a draft "**Model Card**" or technical document of the model while developing. Include: purpose of the model; algorithm used; training data (source, dates, representativeness); performance metrics (overall and by subgroup); known limitations (e.g. "not tested for children under 18"); supposed appropriate and inappropriate uses. This document can be refined and published internally (or externally if appropriate) during the implementation phase.
- **Peer Review and Release Gate:** Before "approving" a model for deployment, have an **independent technical review** - another data scientist or a group (perhaps a *Model Review Committee*), who have not directly participated in the development, review the model's code, the validation results and whether all previous governance steps have been met. They can use a checklist: data sources approved, bias testing done, explanations generated, etc. This gate helps to catch anything that the team may have overlooked. Only after correcting any notes does the model get the green light.
- **Model security:** If relevant, test for security vulnerabilities: for example, carry out *adversarial* proof-of-concept *attacks* (perturbed input) to see if the model is easily fooled; try to extract data by inverting the model (to see if it memorizes personal data); run *fuzzing* (random inputs) and see how it behaves. If weaknesses are identified, apply countermeasures or note in the documentation that the model should not be exposed in a certain way.
- **Deployment and Monitoring Planning:** During development, the team should prepare a **deployment and monitoring plan** for the next phase -

defining performance metrics in production (KPIs), how often the model will be re-evaluated, an update strategy (will it be retrained periodically? under what conditions?), fallback plans if the model becomes unavailable or needs to be shut down. This shows proactive governance: don't launch and forget, but already define a continuous life cycle.

And it's vital to **involve the business and compliance parties** at the end of development for a cross-functional *sign-off*. For example, legal reviews whether the conditions of use (terms and privacy) would cover the new model; compliance checks whether, say, a credit model is in line with fair credit regulations (or whether it will need approval from the regulator). This *final documentation and sign-off* closes the development phase.

D) Final Testing, Independent Validation and Approval Phase

Between development and production, many organizations have a **final validation** or *pre-deployment* phase:

- **Controlled Environment Testing (Pilot/PoC):** You can first deploy the model in a test environment close to the real one, or conduct a *pilot* with a limited sample of users. Policies may require high impact AIs to undergo a supervised *Beta Test*. *For example*, a bank could run a credit granting algorithm in parallel to the human process for a few months (without impacting customers) to compare decisions and only then use it "for real". This makes it possible to check performance under real data and adjust before full use.
- **Performance and Robustness Verification:** Confirm that the metrics obtained in development hold up on new test data. If there are large divergences, investigate why (perhaps overfitting, or perhaps production data differs from training data => need to collect more representative data). Only proceed if results are satisfactory and within the acceptance criteria defined in the requirements.
- **Final Legal Compliance:** Before launch, legal/compliance team makes *final check*: if privacy policy needs to be updated, if there is a requirement to notify users or regulators (e.g. certain countries require reporting use of automated decision making and allow opt-outs). If it's a device or app, check that labels, disclaimers or terms of use include the appropriate information (e.g.: "This product uses AI for [purpose]. Decisions can be reviewed upon request.").
- **Go/No-Go approval:** Finally, hold a **go/no-go meeting** with key stakeholders (project manager, AI committee representative, business process owner, etc.). Review everything: results, documentation, plans. If everyone agrees that the model is in line with objectives and principles, approve the move to production. If not, return for adjustments or even

cancel/postpone (in the case of serious problems with no immediate solution).

E) Deployment Phase and Continuous Monitoring

Once in production (making real decisions or interacting with users), AI remains under governance:

- **Continuous Monitoring Policy:** Establish key metrics to monitor and alerts. For example, data and model *drift*: monitor whether input characteristics change significantly over time (e.g. average user profile has changed), and whether performance declines (a sign that it needs retraining or adjustment). Also monitor occurrences of errors or exceptions (e.g. the model rejected too many cases with low confidence), and negative feedback from users. This monitoring can be automated, with people assigned to review reports periodically.
- **Periodic Re-evaluations and Audits:** Policy of **periodically auditing the model** - internally or by a third party - especially for high-risk models. For example, every 6 months re-verify bias and fairness with new production data (as real use can reveal new patterns). Or annually, independent team re-inspects compliance and security. These *check-ups* ensure that the system hasn't gone off track over time.
- **Maintenance and Updating:** Establish a **maintenance schedule**: who is responsible for maintaining the model? Frequency of re-training with latest data (if applicable)? **Upgrade procedures**: if the model is to be replaced by a new version, it must go through the evaluation and testing stages again before *rollout*. You should never silently update a critical model without testing it (just as you wouldn't do with crucial software). Document each update (version v2 of the model entered on such and such a date, with such and such improvements).
- **Monitoring Results and Equity:** In addition to the technical, monitor **real impacts**. For example, if it's a loan model, monitor business metrics (have defaults really fallen?), but also **equity metrics** (has there been a change in the demographic composition of those approved? Is this expected or does it indicate bias?) If unwanted results appear - e.g. the AI approved more loans, but increased defaults - governance triggers the team to refine or reassess whether the objective was correct.
- **Continuous Logging and Documentation:** Keep a **robust log of AI decisions** (where applicable). For example, automated decision systems can store which inputs led to which output and what the end result was. This is necessary for accountability and for future incident investigations. Also, if there is a requirement to report to a regulator (some regulators ask for regular reports from algorithms).

- **Incident Response Plan:** Develop and communicate an **AI incident procedure**. This covers: if a serious error or anomalous behavior is detected (e.g. AI made a harmful wrong decision, or system suffered an adversarial attack), what steps should be taken? For example: notify the governance committee immediately; evaluate pausing or shutting down the system temporarily (perhaps revert to a manual contingency process); investigate root cause; involve communication if incident affects customers or public (transparency about what happened); trigger remediation and learning plan (update processes to prevent repetition). This plan should be aligned with the existing IT/cybersecurity incident response, but adding specific AI layers (such as communicating to those affected if it was a systemic bias).
- **Documentation & Public Model Card:** Possibly publish information externally in line with a commitment to transparency. E.g. a simplified **model card** posted on the website or in response to requests (regulators or professional clients sometimes ask). Also, keep documentation ready if an external auditor (or regulator) requests it.
- **Human-in-the-loop or Continuous Oversight:** Reinforce that, even in production, many systems require human oversight. Ensure that these functions are active: e.g. human moderators reviewing a percentage of algorithmic content decisions to ensure that AI is not deviating from what is expected. Or risk analysts manually checking the borderline cases that the model has marked as uncertain. This continuous supervision serves both as quality control and as a guarantee against progressive misalignment.

F) Deactivation or Replacement Phase

Part of the life cycle is termination. It can happen that an AI system is retired or replaced. Governance should cover:

- **Retirement criteria:** Define factors that indicate that a model is no longer suitable (e.g. performance below threshold, emergence of superior technology, change in legal or business context). When reached, plan replacement.
- **Safe Transition:** If you're switching to another model, run both in parallel for a while to make sure the new one doesn't introduce problems, and *phase out* the old one gradually. Notify stakeholders if the change is noticeable (e.g. "we've updated our algorithm for better accuracy, we expect a better experience").
- **Record retention:** Even after deactivation, retain documentation and logs for the required period (e.g. financial regulations may require records to be kept for X years). This is for future accountability should any retroactive effects arise (e.g. a customer complaining years later).

- **Lessons Learned:** Carry out a *post-mortem* of the model: what worked well? Were there any incidents? What can be learned for future projects? Incorporate these lessons into policies and perhaps train staff on them.

G) Privacy and Security Policy Update for AI

It is specifically mentioned in the BoK: **evaluate and update existing privacy and security policies for AI**. This is important because AI can introduce new scenarios not covered by traditional IT policies:

- An internal privacy policy may need to include guidelines for anonymization of training data, retention periods for modeling data and rights of data subjects in the face of automated decisions (in accordance with data protection laws).
- Information security policies should incorporate **AI threats**: e.g. add adversarial attacks, data poisoning, model theft to the security risk catalog and plan controls (such as restricting access to models, monitoring anomalies in inputs).
- Access control policies: ensuring that only authorized personnel access training data and models (principle of least privilege). The models themselves can be valuable assets - deciding whether they are considered confidential information (e.g. trade secrets) and protected as such.
- Incorporation of **OWASP ML Top 10** or similar into secure development and testing policies. This includes items such as robust input validation for models (avoiding adversarial injection), protection against model extraction (rate limiting in APIs), and logging of suspicious activity. Security teams should familiarize themselves with these specific vectors in order to integrate them into AI-adapted audits and penetration tests.

H) Third Party Risk Management and Supply Chain in AI

Another part of C.I. is **managing third-party risks** in the context of AI. This includes:

- **AI Technology Providers:** If the company buys AI software (e.g. a recommendation engine from a vendor) or uses cloud services (vision APIs, MLaaS), then it should include these providers in the governance process. Demand information from them about their algorithms, adherence to principles (there are cases of ethical due diligence - ask if the vendor has its own responsible AI program). **Contracts** should cover: performance levels, audit rights, liability in the event of a gross error, privacy clauses (e.g. guaranteeing that the AI provider will not reuse our data inappropriately), and security obligations (incident reporting, compliance with standards).

- **Outsourcing Development or Annotation:** If model development or data labeling is outsourced, choose reliable partners and include data quality, confidentiality and even ethical compliance requirements in the contract (e.g. prohibiting the use of child labor in labeling, or requiring them to respect biases). Carry out **quality controls** on labeled data deliveries (sample and check that there is no carelessness or bias introduced by the labelers).
- **Use of Pre-Trained Third-Party Models (Transfer Learning):** If the team imports open-source or pre-trained models (e.g. an open language model), this is a supply chain risk: the model may contain biases or backdoors. Policies should require **validating these models as if they were in-house**, possibly re-training in part with own data to align. And analyze the license (some open models have restrictions on commercial use).
- **Partnerships and External Data Acquisition:** If you depend on data flows from partners (e.g. data sharing agreement), ensure that they maintain collection and quality standards. Incorporate these dependencies into risk management: what if the partner stops providing data or changes their terms? Have alternative plans.
- **Employees and Third Parties (Human Resources):** BoK cites "*human resources*" in the context of third parties - for example, consultants, temps or external employees involved in AI development. Ensure that they undergo the same ethics training and follow the same policies (including NDAs if they handle sensitive data).
- **Monitor Supplier Performance:** Establish performance metrics for AI suppliers and review them periodically. If an AI tool provider has a history of failures or does not meet fairness requirements, reconsider the contract. Integrate AI into corporate *vendor risk management* as well.

Integrating everything: from **birth** (idea) to **operational lifespan** and eventual **retirement**, governance implements a *guardrail* at every stage. This ensures continuous *oversight* and accountability. It also creates a **trail of documentation and justification** - essential if the company needs to demonstrate compliance (for example, for certification or investigation).

In the end, an organization that follows these practices can show that:

- Evaluated risks and needs before using AI.
- Designed with ethics and clear requirements.
- Trained with adequate data and permission.
- It has tested and validated the model through multiple lenses (technical, legal, ethical).
- He consciously approved.
- It continuously monitors and improves, and can respond quickly if something goes wrong.

- It integrates AI into its larger corporate governance structure, leaving no blind spots (such as third parties or pre-existing policies).

That's the essence of "**comprehensive AI governance**": it's not just one document or committee, but a series of **interconnected procedures throughout the lifecycle**. Implementing all of this is challenging but, in proportion to the risks, necessary - especially in contexts of critical use or high impact. Leading companies already follow many of these practices, and emerging regulations are likely to make them mandatory for certain cases (for example, the EU AI Act formalizes pre-use compliance assessments for high-risk systems, registries, etc., very much in line with what has been described).

VIII- Understand organizational approaches to AI governance

It is important to note that there is no single way of organizing and implementing an AI Governance Program that is applicable to all organizations. **The organizational approach to this issue will depend on various factors, such as the sector in which the company operates, the role that Artificial Intelligence plays in the company's operations, its size, its maturity in this area, etc.** In other words, **organizations need to adapt governance to their own reality.**

The following are typical elements of organizational approaches to AI governance:

- **Centralized vs. decentralized structures:** Companies can adopt AI Governance in a centralized, decentralized or a mixed/hybrid model:
 - **CENTRALIZED:** governance is established by a central team or committee (for example, an "AI Governance" department), defining all AI policies, guidelines and standards for the entire organization. This approach provides **consistency and unified control**, but can introduce slow implementation at the edges and overload the central team.
 - **DESCENTRALIZED:** each business unit or product team is responsible for managing AI governance within its scope, establishing its own processes in line with general principles defined by the organization. This model provides teams with agility and less bureaucracy, but runs the risk of generating inconsistency and a lack of coordination, since areas can follow different paths and decisions regarding governance.
 - **MIXED/HYBRIDIZED:** this model is a combination of CENTRALIZED and DE-CENTRALIZED - a central team defines the general governance principles, policies and tools, while the business units

adapt and apply these guidelines in their specific contexts. The aim is to define a balance between consistency (highlighted in the CENTRALIZED model) and flexibility (highlighted in the DECENTRALIZED model).

- **EXAMPLE:** An organization creates a central AI ethics committee that establishes principles and requires *all* areas to carry out algorithmic risk assessments before deploying AI systems; however, **local product teams** are responsible for conducting these risk assessments on a day-to-day basis and implementing the necessary mitigation measures in their projects, with autonomy for operational details.

DOMAIN II

Understand how laws, standards and frameworks apply to AI

IX- Fundamental Privacy and Data Protection principles

Before the introduction of the main current Data Protection laws, such as the GDPR, the CCPA, PIPEDA and the LGPD, there were two historical sets of principles that established ethical and good practice guidelines to guide the responsible processing of personal data.

- **Fair Information Practices:** Corresponds to the fundamental set of guidelines governing the collection, use and protection of personal data. They have served as the **backbone of much of the world's data protection legislation**.

In the 1970s, with the advance of information technology and the increase in automated processing of personal data, the need to establish standards and guidelines to protect individuals against the misuse of their information became evident. In addition, the cross-border flow of data between countries grew, which began to raise concerns.

In this context, FIPs emerged in the 1970s, and were initially referred to in a report called "**Records, Computers and the Rights of Citizens**", published by the US *Department of Health, Education and Welfare* (HEW). The general principles, translated into **recommendations and guidelines**, generally include:

1. **Access and Amendment.** Agencies should provide individuals with appropriate access to PII and appropriate opportunity to correct or amend PII.
2. **Accountability.** Agencies should be accountable for complying with these principles and applicable privacy requirements, and should appropriately monitor, audit, and document compliance. Agencies should also clearly define the roles and responsibilities with respect to PII for all employees and contractors and should provide appropriate training to all employees and contractors who have access to PII.
3. **Authority.** Agencies should only create, collect, use, process, store, maintain, disseminate, or disclose PII if they have authority to do so, and should identify this authority in the appropriate notice.
4. **Minimization.** Agencies should only create, collect, use, process, store, maintain, disseminate, or disclose PII that is directly relevant and necessary to accomplish a legally authorized purpose, and should only maintain PII for as long as is necessary to accomplish the purpose.
5. **Quality and Integrity.** Agencies should create, collect, use, process, store, maintain, disseminate, or disclose PII with such accuracy, relevance, timeliness, and completeness as is reasonably necessary to ensure fairness to the individual.
6. **Individual Participation.** Agencies should involve the individual in the process of using PII and, to the extent practicable, seek individual consent for the creation, collection, use, processing, storage, maintenance, dissemination, or disclosure of

PII. Agencies should also establish procedures to receive and address individuals' privacy-related complaints and inquiries.

7. **Purpose Specification and Use Limitation.** Agencies should provide notice of the specific purpose for which PII is collected and should only use, process, store, maintain, disseminate, or disclose PII for a purpose that is explained in the notice and is compatible with the purpose for which the PII was collected, or that is otherwise legally authorized.
8. **Security.** Agencies should establish administrative, technical, and physical safeguards to protect PII commensurate with the risk and magnitude of the harm that would result from its unauthorized access, use, modification, loss, destruction, dissemination, or disclosure.
9. **Transparency.** Agencies should be transparent about information policies and practices with respect to PII, and should provide clear and accessible notice regarding creation, collection, use, processing, storage, maintenance, dissemination, and disclosure of PII.

FAIR INFORMATION PRACTICE PRINCIPLES



Later, in 1980, the OECD refined these recommendations in its "**Guidelines Governing the Protection of Privacy and Transborder Flows of Personal Data**", seeking to create a consensus among member countries to deal with the emerging challenges of the information age. The document was not a binding treaty, but rather **recommendations to member countries to adopt legislation and practices that comply with these principles**. The principles are practically the same as those set out in the American report of 1970.

In the same context, it was also recommended that member countries should not restrict the international transfer of personal data to other countries that also comply with the guidelines or have adequate data protection safeguards in place.

Later, in 2013, these principles were updated by the OECD itself to reflect the new challenges in a more interconnected era.

- **Privacy by Design:** Developed in 1995 by Ann Cavoukian, who at the time was the Information and Privacy Commissioner for Ontario, Canada.

The concept of *Privacy by Design* represents a relevant evolution, as it addresses Privacy and Data Protection not just as a patch or secondary concern, but with a proactive approach that must be applied as a central integrated element from the earliest stages of design and development.

The established principles can be summarized as follows:

1. **Proactive, not reactive:** The aim is to anticipate and prevent privacy problems before they occur, rather than reacting to incidents after they have occurred;
2. **Privacy by Default:** By default, the settings of a system or service should offer the highest level of privacy to the user, without the need for any individual action to activate it.
3. **Privacy Embedded into Design:** Privacy should be an integral part of systems architecture and business practices, not an optional add-on or feature.
4. **Full Functionality - Positive Sum, not Zero-Sum:** The goal here is to achieve not only the organization's objectives, but also to combine them with privacy requirements. It is not hoped to create false dichotomies of *privacy vs. security* or *privacy vs. functionality*. The aim here is to achieve both.
5. **End to end security:** Robust security is essential for privacy. Data protection must cover the entire life cycle of information, from collection to destruction/secure disposal.
6. **Visibility and Transparency:** Data collection and use practices must be transparent to users and regulatory authorities. Compliance must be verifiable.
7. **Respect for user privacy:** Everything should be centered on the user - the design should prioritize their interests, offering measures such as appropriate warnings and user-friendly options for managing their privacy.

Ann Cavoukian also referenced the FIPs, but chose to establish guidelines more linked to practice, propagating a methodology to incorporate these principles from conception, right from the start of the process.

X- AI laws and regulations

Advanced policy debates on AI risk emerged only recently.

Thus, unlike the current broad framework in relation to Privacy and Data Protection, with many countries having specific legislation on the matter, the regulation of Artificial Intelligence is still growing and taking shape. To give you an idea, the EU AI ACT, which came into force in August 2024, is considered the world's first comprehensive legislation designed specifically to regulate the use of AI systems in EU member states.

With this in mind, as AI becomes ubiquitous and specific regulations are still being discussed, various "parallel" laws have been and are still being used to establish the ethical and responsible use of technology.

- **Data Protection**

The GDPR is considered to be the benchmark global legislation on the subject of Data Protection and Privacy, having inspired several other pieces of legislation around the world, including Brazil's own text, the LGPD - General Data Protection Law.

In force since May 2018, the GDPR imposes strict obligations on the use of personal data, establishing principles (including in line with the FIPs) and rights for data subjects.

In the context of AI, as it is understood that many systems process personal data, it is also applicable that companies using such systems are also responsible for applying and complying with the rules established by the GDPR, if applicable to the hypotheses of legal application of the standard.

EXAMPLE: A hospital decides to implement an AI system to analyze patients' medical data, which is considered sensitive personal data. In this way, the hospital must comply with the basic principles of the GDPR, such as transparency, minimization and security, assess the applicable legal basis for the use of this data, establish transparency for patients about the use of their data and allow, where applicable, the regular exercise of their rights, free of charge and in an accessible manner.

In addition to the GDPR, we can also include other legislation in our comments, from Canada's PIPEDA (Personal Information Protection and Electronic Documents Act), Australia's Privacy Act, Chile's new data protection law and the LGPD - General Data Protection Law, enacted in 2018.

Despite local peculiarities, it is clear that, in many contexts, the Data Protection and Privacy regulatory model for various countries follows a similar model, establishing guiding principles, rights for data subjects, obligations for organizations, the existence and supervision of supervisory authorities/agencies, the need to report security incidents and other normative characteristics that are considered "standardized" in an international scenario under analysis of the matter.

As a result, both Deployers and Providers, when using AI systems, must also follow local Data Protection and Privacy regulations, applying the necessary principles and analyzes.

- **Intellectual Property**

Intellectual Property is another essential legal field to study in relation to Artificial Intelligence, as it deals with patents, copyrights, trademarks and trade secrets. The rise of AI has brought new challenges as to who owns the rights to machine-generated creations and how existing laws apply.

A) **PATENTS AND INVENTORSHIP:** Patent laws traditionally **require a human inventor.**

In this context, we can cite a recent case in the USA (*Thale v. Vidal*), which takes place in 2023. In 2019, Stephen Thaler submitted two patents to the USPTO listing only his AI system "DABUS" as the inventor. The USPTO rejected the application, claiming that the Patent Act states that the inventor must be a "natural person". Following the decision, Thaler appealed. The District Court of Virginia and the Federal Circuit upheld the USPTO's interpretation. The US Supreme Court refused to consider the issue. Thus, currently, considering the case law, only patents created by **HUMAN BEINGS / NATURAL PERSONS** are allowed to be registered in the US, **although the agency has already issued specific guidelines for AI "assisted" inventions in 2024.**

In other countries, such as Europe, patent agencies have also rejected applications with AI as the inventor, stressing that the invention requires **human legal personality.**

B) **COPYRIGHT AND WORKS GENERATED BY AI:** This is undoubtedly one of the main issues of conflict related to Artificial Intelligence.

Copyright aims to **protect original works of authorship fixed in a tangible medium (text, images, music, etc.).** Thus, a relevant question arises: **can content created by AI (such as a generated image) be protected by Copyright?**

In several countries, the answer is no, **because there is no human author.** The US Copyright Office has been denying registration to works purely generated by Artificial Intelligence without human creative involvement.

In addition to this discussion, [there are extensive debates regarding the use of copyrighted materials in training Generative AI models.](#) There are several relevant real-life cases taking this discussion as their object:

❖ **[SILVERMAN v. OPENAI \(2023\)](#)**

A group of authors, such as Sarah Silverman, Christopher Golden and Richard Kadrey, have filed a class action lawsuit against OpenAI, the developer of ChatGPT, claiming that OpenAI has used their books, obtained from pirated repositories, to train GPT models, in violation of copyright and the California Unfair Competition Law and the DMCA (Digital Millennium Copyright Act). Currently, OpenAI is facing other similar lawsuits in the US and, as a result, several are being evaluated together in the Southern District of New York, under MDL No. 3143, assigning Judge Sidney H. Stein to conduct the case and suspending the hearings in California.

In February 2024, Judge Araceli Martinez-Olguín upheld the central charge of direct infringement (and a limited version of the unfair competition claim), but rejected the theses

of vicarious/contributory infringement (in which the plaintiffs failed to show substantial similarities between GPT's outputs and their books) and related to the DMCA, due to the absence of facts about intentional removal of CMI.

❖ The New York Times Company v. Microsoft Corp & OpenAI

In December 2023, The New York Times filed a lawsuit in New York alleging that Microsoft Corp and OpenAI had used **millions of paywall articles to train ChatGPT and Bing AI, violating copyright and causing a loss of subscriptions and advertising revenue.**

In March 2025, Judge Sidney H. Stein rejected most of the defendants' motion to dismiss, keeping alive the charges of direct and contributory infringement, rejecting the statute of limitations argument. This is one of the cases centered in Stein's hands, as indicated in SILVERMAN v. OPENAI.

In both cases, the main legal debate focuses on the extent to which *Fair Use*, argued by the big-tech companies and considered an exception to the application of US Copyright Law, would also apply to this case.

The big-tech companies compare the analysis mainly in a relevant case law involving [Google Books - in 2005, a company sued Google for digitizing millions of books for its search engine without authorization from the rights holders.](#) Google claimed that the copies served merely to index and display brief "previews", and did not make complete works available to the public. On November 14, 2013, there was a summary decision in favor of Google, declaring that the Google Books project fell under Fair Use, classifying it as "highly transformative", being beneficial for research and preservation, and harmless to the publishing market. The Second Circuit, a year later, confirmed the understanding, reinforcing the four factors that favored Google in the use of Fair Use. The Supreme Court refused to consider the Authors Guild's appeal, ending the case and consolidating the understanding that the digitization and display of previews by Google Books constitute legitimate use under US law.

- C) **TRADEMARKS AND TRADE SECRETS:** AI can both infringe trademarks (for example, generate logos or names similar to existing and duly registered trademarks) and help protect them (there are already systems in place to detect the misuse of trademarks by Artificial Intelligence).

In addition, a relevant challenge here is precisely what is meant by **trade secrets**. One challenge is that AI models, such as *Deep Learning*, are often associated with "black boxes", the exact *know-how* learned being difficult to delineate, which makes protection complex.

Moreover, if employees or competitors can gain unauthorized access to a particular company's proprietary models or data, this could also constitute misuse of trade secrets. In practice, there is latent protection within organizations regarding the use of AI models and personal and sensitive/confidential information/data, with many organizations adopting policies restricting the use of certain data associated with these AI models or even restricting use internally, vis-à-vis their employees.

In this sense, a relevant piece of news was that, according to Bloomberg, [in 2023 Samsung banned the use of ChatGPT among employees after the accidental leak of an internal source code by an engineer who sent it to ChatGPT last month](#). After learning about the incident, the South Korean company issued a memo banning the use of GenAI tools, showing the company's zeal for the data shared by its employees with Generative AI models.

- **Health**

In the US, the **AFFORDABLE CARE ACT (ACA)**, enacted in 2010, more specifically the interpretation given to **Section 1557**, has become one of the main federal instruments to combat algorithmic *bias* in US healthcare.

In 2024, HHS/OCR expanded the scope of the rule, making it clear that *patient-care decision-support tools* - including *machine learning* algorithms purchased from third parties - cannot discriminate against patients or policyholders on the basis of race, color, national origin, sex, disability or age. The rule requires **hospitals, health plans and insurers to inventory, test for and mitigate any biases, under penalty of losing federal funding or facing sanctions**.

The rule began to apply to algorithms implemented as of July 5, 2024 and to legacy systems as of May 1, 2025 in the US.

In this context, below are the main practical obligations for entities covered by the ACA:

- a- **Inventory and classification:** Organizations must map each algorithm (CDS, screening, claims, utilization management) that could affect protected groups.
- b- **Bias testing:** Auditing performance stratified by race, gender, age and disability: comparing error rates and impacts.
- c- **Mitigation and recording:** Re-train, re-calibrate or discontinue problematic models: document the entire correction cycle for OCR inspiration.
- d- **Integrated governance:** Align these steps with frameworks such as **NIST AI RMF** and the value requirements in *Medicare / Medicaid* programs.
- e- **Transparency for patients:** Update non-discrimination notices and create automated decision review channel.

- **Fairness, Non-Discrimination and Consumer Protection**

One of the main concerns about the use of AI systems in everyday business processes is the risk of discrimination. Thus, with the intensification of the use of technology in recent years, specific legislation has emerged to mitigate and establish controls on discrimination in selection processes, credit analysis, housing, immigration, among other impact scenarios.

In the US, for example, the **EEOC (Equal Employment Opportunity Commission)** has issued guidelines on the use of Artificial Intelligence in automated hiring tools, which must comply with employment equality laws and that employers can be held liable if software carries out a pre-selection of candidates that could be considered discriminatory, such as evaluating candidates by assessing skin color, religion, sex and/or nationality.

In New **York**, there is **New York Local Law 144**, specifically designed to regulate the use of algorithmic tools in employment-related decisions - **AEDT - Automated Employment Decision Tool**. This law, which came into force in 2023, requires companies that use AI in hiring processes to carry out annual independent audits to assess and identify possible discriminatory biases before use and to notify the candidate when a decision has been automated.

The state of Illinois (USA) has enacted the **AI Video Interview Act**, which obliges employers to inform candidates when AI is used to analyze recorded interviews and delete the videos upon request.

In the context of consumer protection, organizations such as the Federal Trade Commission (FTC) in the US and other equivalent bodies at an international level can impose penalties for algorithmic practices that are considered unfair or deceptive. For example, if a virtual assistant deliberately omits options or a recommendation system manipulates children into *in-app* spending, such conduct can be investigated under consumer protection laws.

- **Civil (and criminal) liability for algorithmic decisions**

To discuss this topic, the following question is valid: **if an autonomous vehicle causes an accident, who should be held accountable - the manufacturer, the developer of the vehicle's autonomous driving model, the "driver" present in the vehicle or the vehicle itself (in this case, the AI)?**

Currently, civil liability laws apply to AI indirectly, via theories that have already been consolidated: **negligence, liability for defective products, professional error, etc.**

Likewise, it is also necessary to assess the cause and the link - in any incidents involving AI, technical investigations must be carried out to identify whether the error was the system's or the person operating it.

In the EU, for example, in September 2022, the European Commission released the proposed **AI Liability Directive**, which deals with claims for damage caused by AI systems or the use of AI, adapting the rules of non-contractual civil liability to AI. In this way, the standard would complement the EU AI ACT by introducing a new liability regime, with the aim of establishing greater legal certainty, increasing consumer/user confidence in the use of technology and helping with liability claims for damage caused.

In any case, the proposal was shelved and the European Commission has no plans to resume debate on the bill, citing a lack of agreement, as the technology industry has been pushing for simpler regulations that promote innovation. In this same context, it is also important to note the harsh criticism made by US Vice-President J.D Vance at the AI Action Summit promoted by the French government in Paris in February 2025, who called for

European countries to embrace the "new frontier of AI with optimism rather than apprehension", aiming for a lighter approach to regulating the technology.

Specifically with regard to criminal liability, there is no specific information on whether the AI system itself has been indicted or held directly responsible, but there is an investigation into the human beings behind the system. Debates about giving AI autonomous legal personality have arisen academically, but have not yet been adopted in legal frameworks, and it is a more accepted trend to frame AI as a tool for which humans (developers, operators and companies) are accountable.

- **California laws**

In addition to the indirect application of the CCPA (California Consumer Privacy Act) for issues related to privacy and data protection, California also has regulations that aim to establish obligations and guidelines related to the governance and implementation/use of AI systems.

(I) **GENERATIVE AI TRAINING DATA TRANSPARENCY ACT - AB 2013**

The central objective is to establish public transparency of the datasets used to train any *GenAI* or AI services made available to California residents. The scope of the law is generative AI systems launched and/or updated on or after 01/01/2022, and public or private developers who develop free or paid AI systems for the Californian public.

By 01/01/2026, developers must publish detailed documentation on their website:

- Summary of each *dataset*;
- Origin, purpose, quantity and type of data (including whether there is personal data or material protected by *copyright, trademark or patent*);
- Use of synthetic data;
- Collection period and date of first use;

As for penalties, there is no indication of a fixed amount, but it does authorize civil action by the **Attorney General, including the possibility of preventing the service/system from being offered.**

The requirement to disclose data sources allows for external audits of biases, representativeness and potential discrimination built into the training of models. Developers will need to assess and justify the appropriate diversity/quality of data to avoid litigation.

(II) **CALIFORNIA AI TRANSPARENCY ACT - SB 942**

The central aim of the regulation is to **guarantee the provenance and labeling of AI-generated content (text, image, audio and video).**

The regulation applies to any person or entity that develops a GenAI model with more than 1 million monthly visitors and that is publicly accessible in California.

From 2026, organizations must:

- Provide the user with a **free AI detection tool** to check if content has been created/changed by the system;
- Offer the option of **manifest disclosure** (visible mark) and oblige **latent disclosure** (permanent metadata) that identifies that the content is generated by AI and the name of the provider/time and unique identifier.
- Revoke license within 96 hours if the third party removes the **disclosure** capability.

As for sanctions, a civil fine of USD 5,000 per day is applicable for **Covered Providers**, in addition to the possibility of injunctions and costs against third parties.

Although the focus of the law is **content transparency**, the law facilitates the detection of discriminatory and disinformative *deepfakes* by creating technical mechanisms that support accountability for biased use.

(III) **BOLSTERING ONLINE TRANSPARENCY (B.O.T) - SB 1001**

The central aim of the regulation is to **ban the use of unidentified bots posing as humans to influence elections or induce purchases**. The law came into force in 2019.

The scope of the law is any "bot" operating online, performing mostly automated interactions and applicable to websites/applications with more than 10 million unique users in the US in the last 12 months.

There are no set amounts for non-compliance, but violations can be prosecuted by the **Attorney General** as a deceptive practice.

• **IAI ACT**

The European Union has established itself at the forefront of regulation on the subject with the introduction of the EU AI ACT, the world's first horizontal regime dedicated to AI.

- A) **MATERIAL AND TERRITORIAL SCOPE:** The regulation regulates AI systems, prohibited or high-risk AI practices, general purpose AI models (GPAI) and transparency obligations. In addition, it may apply outside the territorial limits of the EU, whenever the system or its result may reach European territory.
- B) **OPERATORS:** The EU AI ACT defines specific agents that relate distinctly to the AI system:
 - a. **PROVIDER (Developer):** is the entity that develops an AI system or a GPAI and puts it on the market under its own brand.
 - i. **EXAMPLE:** ChatGPT, Mistral, DeepSeek, Anthropic

- b. **DEPLOYER:** A natural or legal person who uses an AI system within their professional activities, under their authority (not to be confused with the lay end user).
 - i. **EXAMPLE:** A European bank acquires a license for an AI system that performs initial screening of CVs, provided by a third-party company (PROVIDER).
- c. **DISTRIBUTOR:** A distributor resells or makes available the AI system within the EU without altering its properties.
 - i. **EXAMPLE:** A Danish reseller that distributes third-party APIs in several European countries.
- d. **AUTHORIZED REPRESENTATIVE:** A person or company established in the EU that receives a written mandate from a PROVIDER located outside the bloc to represent it before authorities, keep technical documentation available and cooperate with any enforcement actions.
 - i. **EXAMPLE:** An international law firm based in Brussels acts as the official representative of a Chinese AI provider focused on medical care.

C) **RISK STRUCTURE:** EU AI ACT takes a risk-based approach, defining graduated categories:

- a. **UNACCEPTABLE RISK (Prohibited AI Practices):** AI systems whose use and development is prohibited;
- b. **HIGH RISK (High Risk AI Systems):** Subject to stricter requirements before and after marketing and making available to the public;
- c. **LIMITED RISK:** Obligations related to transparency and other specific control measures;
- d. **MINIMUM RISK:** Free use.

D) **PROHIBITED IA PRACTICES**

- a. Subliminal or manipulative techniques that distort behavior or cause significant harm;
- b. Exploitation of possible vulnerabilities, such as age, disability or socio-economic status;
- c. Social Scoring that generates unjustified unfavorable treatment;

- d. Prediction of crime based solely on profiles or personal characteristics;
- e. Construction of facial bases by indiscriminate scrapping;
- f. Inference of emotions in a work or educational environment (except for medical and security purposes);
- g. Biometric categorization to infer race, sexual orientation, religion and other intimate aspects of a person;
- h. Remote biometric identification in real time, in places accessible to the public, for law enforcement, except in strictly authorized exceptions.

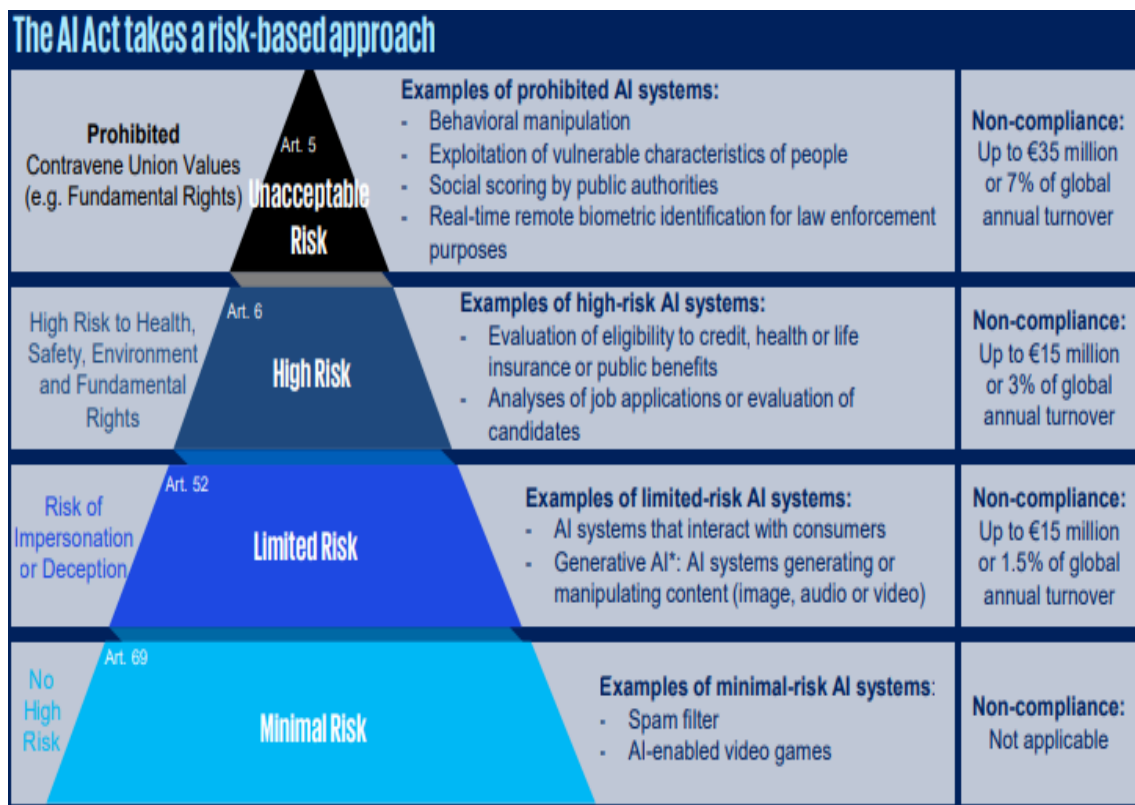
E) HIGH-RISK AI SYSTEMS

- a. **Regulated products: health** sector systems (algorithm that evaluates x-ray images for diagnosis); **traffic** (highway autopilot in cars); **infrastructure and gas** (AI-powered safety valve controller); **aviation** (autonomous landing support system) and **toys** (educational robot that avoids collisions and promotes learning); **elevators**.
- b. **Biometrics:** remote biometric identification in real time (biometrics in soccer stadiums), biometric categorization or *emotion recognition*.
- c. **Critical infrastructure:** protection of electrical networks, *data centers*, cloud, water, gas and telecom;
- d. **Education and training:** selection of students in selection processes, exam and classroom surveillance, automated correction of tests and quizzes.
- e. **Employment and employee management:** CV screening, performance evaluation, dismissal of employees.
- f. **Essential services and benefits:** granting credit, public benefits, screening emergency calls or public assistance.
- g. **Law Enforcement:** crime risk definition, recidivism analysis, evidence evaluation;
- h. **Migration and border control:** visa screening, document fraud detection, automated polygraph;
- i. **Justice and demographic processes:** support for the establishment of court rulings, alternative dispute resolution, political micro-targeting.

Furthermore, exceptions may exist if all of the following conditions are met. Furthermore, it is PROVIDER's responsibility to document the analysis, register its AI in the EU base and bear the burden of proof.

❖ **NON-SIGNIFICANT RISK TO HEALTH, SAFETY OR FUNDAMENTAL RIGHTS, where the purpose falls under at least one of the four scenarios:**

- **Narrow procedural task:** for example, a tool that converts unstructured PDF into a spreadsheet;
- **Improving completed human activity:** AI that adjusts the professional tone of a **finished** text.
- **Detecting patterns or anomalies without replacing human decision-making:** a system that identifies incoherent school grades for manual review by a teacher.
- **Preparatory task:** Automatic document translator before legal analysis.



F) OBLIGATIONS FOR HIGH-RISK AI SYSTEMS

Considering the risks applicable to the development and implementation of high-risk AI systems, agents must adopt certain measures:

➤ **Design and development**

- **Pre-registration in the EU Database:** The Developer (Developer / Provider) must enter the system's metadata, for public transparency purposes, in the EU Database, which is a catalog of AI systems.
- **Risk-Management System:** the Developer / Provider must implement an ongoing process to identify, analyze, evaluate and mitigate risks to health, safety or fundamental rights. All this must be re-assessed with each relevant change.
- **Data Governance:** The Developer must use training/validation/test sets that are relevant, representative and free of serious errors and without undue bias. In doing so, they must document the origin, composition and preparation of the *dataset*.
- **Technical Documentation:** Produce a complete dossier: purpose, architecture, hardware, data requirements, performance metrics, tests, validations and model versions.
- **Automatic Record Keeping:** Enable the automatic recording of relevant events during the lifecycle, with sufficient granularity for auditing and investigating any incidents.
- **Human oversight:** Design the system to allow intervention by qualified people, define criteria, alerts and tools that maintain effective human control.
- **Accuracy, robustness & cybersecurity:** Achieving and declaring levels of accuracy consistent with the intended use and ensuring resilience against failures and adversarial attacks.

➤ **Go-live**

- **Transparency and instructions:** Provide manual to Deployer containing: capabilities, limitations, accuracy metrics, data requirements, how to interpret output, security measures and maintenance.
- **Fundamental Rights Impact Assessment (FRIA):** The Deployer, whether public bodies or private entities, must carry out the FRIA **BEFORE THE FIRST USE OF THE SYSTEM**.
- **Activate human supervision logs and controls in the live environment:** Providers and Deployers must activate human supervision control functionalities and logs for monitoring and auditing purposes.

➤ **After-Market (Continuous Operation)**

- **Post-Market Monitoring Plan:** Collect actual performance, feed back into the QMS (**QUALITY MANAGEMENT SYSTEM**) and correct deviations and emerging risks.
- **Serious-Incident Reporting:** Notify the surveillance authority in less than 15 days (or less than 2 days if the risk is very serious).
- **Update the QMS (QUALITY MANAGEMENT SYSTEM), Risk Management and documentation:** Whenever there is a substantial change or incident.
- **Cooperation with authorities:** Keep documents available and respond to requests.

G) Supervision

For the entire European Union, the **European AI Office** will be the authority responsible for the coordination and uniform application of the EU AI ACT. Compared to the Privacy Framework, it resembles the EDPB - European Data Protection Board.

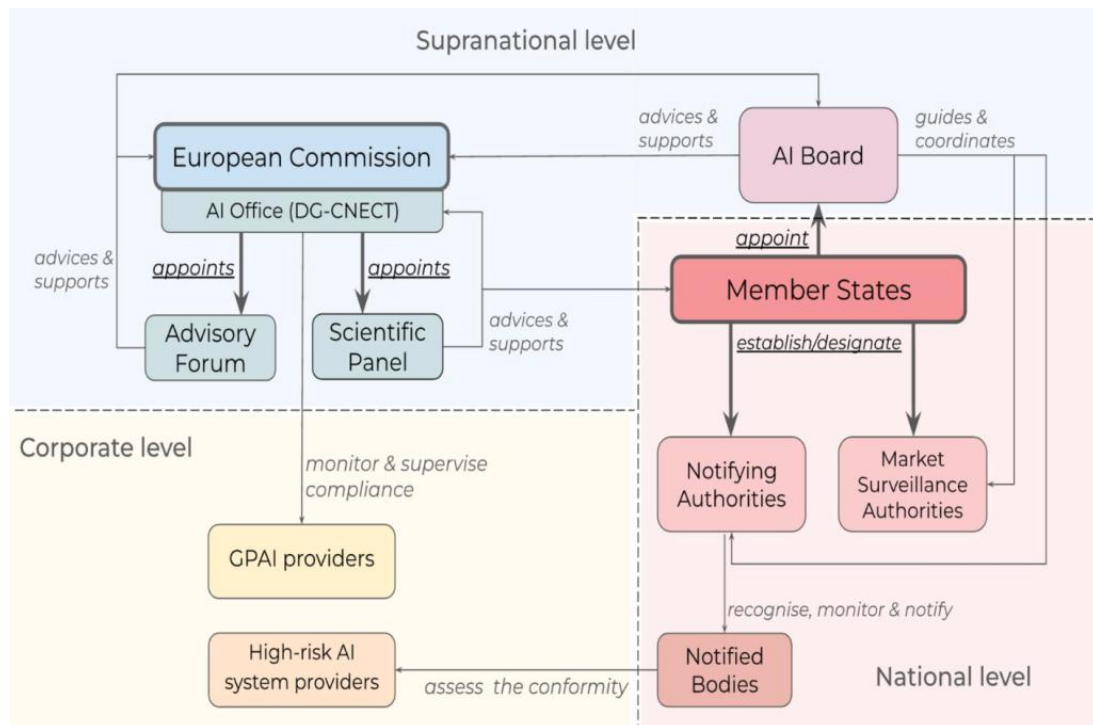
The **EU AI Office** directly supervises **GENERAL-PURPOSE AI MODELS (GPAI)**, especially those with *systemic risk*, and can request technical audits, order risk assessments and apply corrective measures. It is also responsible for managing the public database of high-risk systems (EU Database) and for drawing up **codes of good practice and guidelines** to help agents and other local authorities apply the issue.

For the context of the EU AI ACT, GPAI with high impact potential have 10^{25} Flops.

In addition to the EU AI Office, we have the **EUROPEAN AI BOARD**, made up of a representative from each member state and the EU AI Office's own commission. It issues opinions, resolves disagreements between local authorities and can define good practices and recommendations.

At the national level, each member state will be able to establish **THREE TYPES OF AUTHORITIES** as part of the implementation of the EU AI Act.

- 1- **Market Surveillance Authority:** must carry out activities and take measures relating to market surveillance and product compliance. In general, e;a will have the task of ensuring that only products that comply with European laws are made available.
- 2- **Notifying Authority:** responsible for establishing and implementing the procedures for assessing, designating and notifying conformity assessment bodies and for monitoring them.
- 3- **National Public Authority:** member states must identify their National Public Authorities that enforce the obligation to respect fundamental rights in relation to high-risk systems, with powers to request or access any documentation created or stored in accordance with the EU AI ACT, when such documentation is necessary for the effective fulfillment of their mandate, within legal limits.



Each country is entitled to define its own regulatory scope applicable to the subject. Spain, for example, has defined just one authority - the **Spanish Agency for the Supervision of Artificial Intelligence (AESIA)**, which acts as a single market surveillance body under the Spanish Department for Digital Transformation. Finland, meanwhile, has appointed 10 existing market surveillance authorities, including the Energy Authority, the Transport and Communications Agency and the Medicines Agency.

In the Netherlands, supervision and monitoring will be the responsibility of the **Dutch Digital Infrastructure Authority (RDI)** and the **Dutch Data Protection Authority (AP)**. The supervision of high-risk systems will be the responsibility of the AP and the RDI. The AP will also be responsible for overseeing transparency obligations. The RDI will be responsible for coordination and support on technical aspects.

H) Sanctions

Like the GDPR, the EU AI ACT also provides for administrative sanctions in the event of infringements.

- 4- **Adoption and/or development of AI systems of prohibited practices:** EUR 35 million or 7% of global turnover, whichever is higher.
- 5- **Failure to comply with general obligations** (high-risk requirements, GPAI rules, transparency, etc.): EUR 20 million or 4% of global turnover.
- 6- **Providing incorrect or misleading information to the authorities:** EUR 7.5 million or 1% of global turnover.

- 7- **EU public sector:** Up to EUR 1.5 million for prohibited practices and EUR 750k for other violations.



In addition to fines, the authorities can order technical corrections or adaptations to the model; temporary or permanent bans on use; demand product recalls and publish the identity of the offender - the effect of **naming and shaming (similar to what happens with the LGPD with the sanction of publicization)**.

For startups and SMEs, the EU AI Act allows for a reduction in the amount of sanctions (Recital 103), while maintaining the percentage for large groups.

The sanctions relating to prohibited practices and AI Literacy requirements have been active since February 2025 in the EU, and the remaining sanctions, applicable to obligations on GPAI, EU AI Office duties and all high-risk requirements will apply from August 2025.

- **Other local regulations**

In addition to the EU AI Act, which established the global basis for AI governance, there are other regulations being developed around the world that are worth mentioning:

- 1- **Canada - Artificial Intelligence and Data Act (AIDA):** Inserted in Bill C-27, AIDA establishes a federal "obligations-based compliance" regime for "high risk systems", requiring risk assessments, documentation and internal audits. It also makes it possible to impose fines of up to CAD 25m or 5% of the company's global revenue.
- 2- **Brazil - PL 2.338:** In December last year, PL 2.338/2023 was approved in the Chamber of Deputies after several new texts. Currently, the Bill will begin its discussion in the Chamber of Deputies, which set up a Special Committee in May, chaired by Deputy Luisa Canziani. The text also distinguishes between excessive risk systems (prohibited practices) and high risk systems. It establishes the SIA - National Artificial Intelligence System, which will have a Coordinating Authority (probably the ANPD - National Data Protection Authority) and other sectoral authorities (BACEN, ANS, ANATEL etc).

3- United States - Executive Order 14110 + Colorado SB 24-205 / California SB 1047

- a. **Executive Order 14110 - Safe, Secure and Trustworthy Development and Use of AI (10/30/2023):** The Joe Biden Administration, in alignment with the main big-tech companies, established pillars and guidelines for the safe, responsible and conscientious development of Artificial Intelligence. Despite not being a federal law, Executive Order 14110 marked a commitment by the main Providers on the subject.

The text was repealed on January 20, 2025 by the Donald Trump administration, which in turn issued another Executive Order - *Removing Barriers to American Leadership in Artificial Intelligence*. This was a clear backroom deal between the new Trump administration and Big Tech.

The main points of Executive Order 14110 were:

- i. **Safety and Security:** Focus on ensuring the safety of AI systems by requiring developers of powerful models to share safety test results with the government. Establishes safety standards, addresses risks of AI-enabled hazardous substances and considers watermarking for synthetically generated content. It also creates the Artificial Intelligence Safety and Security Board.
- ii. **Promoting innovation and competition:** Boosting AI research and development in the US, including through the National AI Research Resource and support for small developers (startups). Facilitates the attraction of foreign AI talent by simplifying bureaucratic visa processes. Promotes public-private cooperation and clarity on issues related to Intellectual Property.
- iii. **Support for workers:** Aims to mitigate the negative impacts of AI on the labour market and maximize its benefits for workers. It mandates the creation of reports on the impact of AI on the labor market and the development of principles to protect the well-being of employees.
- iv. **Advances in fairness and civil rights:** Seeks to ensure the non-discriminatory use of AI by combating algorithmic discrimination in the legal and criminal system. Requires the issuance of guidance to federal agencies on discrimination and privacy violations arising from the use of AI. Promotes the responsible development of AI in health, human services, transportation and education sectors.
- v. **Privacy:** Aims to mitigate the privacy risks associated with AI by promoting the development of privacy-enhancing technologies

(PETs) and formulating guidelines on the collection and use of personal data and privacy assessments by the government.

- vi. **Advancing the use of AI in the Federal Government:** Guides the acquisition of AI by government agencies, with each agency designating its **Chief AI Officer (CAIO)**. It also promotes a risk-based approach.
 - vii. **Strengthening American leadership:** Aims to position the US as a global leader in AI through international cooperation and the establishment of a global framework to manage the risks and benefits of the technology.
- b. **Trump Executive Order - Removing Barriers to American Leadership in Artificial Intelligence:** Published by Donald Trump days after taking office in 2025, the document seeks to instruct federal agencies to review existing regulations and eliminate **unnecessary** barriers **to AI innovation** and promote US national competitiveness and security.

Main points:

- i. **Leadership action plan:** The order requires the preparation of an "action plan" to achieve the established policy of **maintaining US leadership in AI**. The plan must be developed within 180 days.
 - ii. **Review and revocation of policies:** The order directs the review of all policies, guidelines, regulations and other actions taken pursuant to Executive Order 14110. Any actions that are found to be inconsistent with or present obstacles to the AI incentive policy should be suspended, revised or rescinded.
- c. **California SB 1047:** Senate Bill (SB 1047), known as the *Safe and Secure Innovation for Frontier Artificial Intelligence Models Act*, was a bill passed by the California legislature with the aim of regulating advanced AI models.

However, on September 29, 2024, California Governor Gavin Newsom, amid much criticism from Big Tech, vetoed SB 1047, despite recognizing the importance of AI safety and the need to adopt safeguards. Newsom pointed out that the bill's approach was not based on an empirical analysis of AI systems and capabilities, focusing more on the cost and size of the models than on their function or potential risks.

Below are the main points of SB 1047 if it had passed:

- i. Publication and implementation of security protocols;
- ii. Adoption of reasonable care to prevent misuse and avoid causing or allowing catastrophic damage, such as the creation of chemical or

- biological weapons, cyber attacks on critical infrastructure or events with many victims;
- iii. Implementation of a "kill switch" to allow the immediate suspension of the AI system at any time in the event of potential damage;
- iv. Carrying out annual third-party audits to assess security controls;
- v. Reporting AI safety incidents to the California Attorney General within 72 hours.

d. **Colorado SB 24-205:** SB 24-205, entitled "**Consumer Protections for Artificial Intelligence**" aims to protect consumers in interactions with AI systems. The law will come into force in February 2026. It establishes obligations for both Providers and Deployers of high-risk AI systems.

- i. **For Providers:** Developers/Providers must take reasonable care to protect consumers from known or reasonably foreseeable risks of algorithmic discrimination.

Obligations of Providers / Developers:

1. Provide a statement disclosing specific information about the high-risk system and specific documentation / information for carrying out an impact assessment (EIA / FRIA);
2. Make public a statement summarizing the types of high-risk systems they have developed and how they manage the risks of algorithmic discrimination;
3. Disclose to the Attorney General and Deployers any known or reasonably foreseeable risks of algorithmic discrimination within 90 days of their discovery.

Obligations:

1. Implement a risk management policy and program for high-risk systems;
2. Carry out a system impact assessment - EIA / FRIA;
3. Annually review the implementation of each system and its use, to ensure that it is not causing algorithmic discrimination;
4. Notify the consumer if the high-risk system makes a consequential decision about them;
5. Warn consumers that they are interacting with an AI system.

6. Provide the consumer with the opportunity to correct incorrect personal data processed by the high-risk system;
 7. Provide the consumer with the opportunity to challenge an adverse consequential decision, through human review if technically feasible;
 8. Make public a statement summarizing the types of high-risk systems they implement and how they manage the risks applicable to algorithmic discrimination.
- 4- **United Kingdom:** The British government has opted to adopt a "light touch" regulatory model, which requires 12 sector regulators to publish plans on how to apply five guiding principles regarding the implementation and use of Artificial Intelligence - **security; transparency; accountability; contestability and proportionality** - in each of their areas of competence.
- 5- **China:** China has **Interim Measures for Generative AI Services**, published in 2023, establishing security assessments, incident reports, human review of the publication of "content that affects public opinion" and joint liability.

Other countries also discussing and evaluating the issue: the United Arab Emirates, Australia, Nigeria, South Korea, Japan and India.

XI- Technical norms and international standards

While laws define what must be done (obligations, prohibitions and rights), technical norms and standards provide **HOW TO - specifications, metrics, practical guidelines for implementing AI in a safe, interoperable and ethical way.**

Several international standardization bodies and international working groups have been working on international AI standards, including ISO/IEC and IEE, as well as other sectoral initiatives.

These guidelines complement legislation - **for example, a law can require an AI system to be "auditable and secure" and a technical standard can offer measurable requirements for auditability and security.**

- **ISO/IEC**

ISO (International Organization for Standardization) and IEC (International Electrotechnical Commission) have formed a joint committee focused on Artificial Intelligence: ISO/IEC JTC 1/SC 42. This committee, created in 2017, has since been working on various international standards for Artificial Intelligence and Big Data. Some of the relevant ISO/IEC standards:

- (i) **ISO/IEC 22989:2022 - Artificial Intelligence:** Concepts and terminology. This establishes a common language of AI terms and definitions, which is important for aligning understandings. For example, there is a definition of what a training *dataset* is, what characterizes machine learning, etc.
- (ii) **ISO/IEC 23053:2022 - Framework for Implementing AI Systems:** describes high-level concepts and steps for implementing an AI system, useful for guiding organizations to structure their internal AI projects.
- (iii) **ISO/IEC TR 24027:2021 - Assessment of bias in datasets:** a technical report that provides guidelines on how to identify and mitigate bias in data used by algorithms, in line with *fairness* concerns.
- (iv) **ISO/IEC 23894:2024 - AI Risk Management:** aims to provide a standardized framework for identifying, assessing and treating AI system risks, possibly aligned with frameworks such as the NIST AI RMF.
- (v) **ISO/IEC 42001:2023 - AI Management System (AIMS):** This standard, divided into parts, provides requirements for organizations to establish an internal management system focused on responsible AI, analogous to what ISO/IEC 27001 does for Information Security. Part 1 establishes high-level requirements (e.g. leadership commitment to ethical AI, policy definition, assessment and continual improvement of AI risks. Part 2 provides specific controls that can be implemented, such as AI ethics committees, algorithmic impact assessment procedures, AI governance training and so on.

In short, ISO/IEC international standards **provide a voluntary technical-normative framework**. Although they are not laws, they can be adopted contractually or referred to by regulators as guidelines or good practices to be followed. For example, the EU AI ACT proposes that the European Commission indicate "harmonized standards" which, if followed, give presumption of conformity with certain requirements of the regulation. In this way, ISO/IEC standards **help organizations stay on top of the state of the art and make it easier to demonstrate regulatory compliance**.

- **IEE (Institute of Electrical and Electronics Engineers)**

Through the IEEE Standards Association, the IEEE has also established initiatives for ethical standards in autonomous and intelligent systems.

Within this scope, the **IEEE 7000 series** family stands out, focused on aspects of human values in the engineering of intelligent systems.

- (i) **IEEE 7000-2021 - Process model for consideration of ethical issues in system design:** provides a formal process for engineers to incorporate ethical values during the design of complex systems, including AI. It helps to map ethical impacts, mitigate dilemmas and document decisions from the start of the project - ***ethically aligned design***.

- (ii) **IEEE 7001 - 2021 - Transparency of Autonomous Systems:** defines measurable degrees of transparency in autonomous systems, including AI. It provides guidelines for designing systems that can explain their actions in a way that is understandable to users and auditors, with levels of transparency adapted to the type of audience - **example:** more technical for regulatory inspection and simpler for lay users. Here we find clear application to the concept of **explainability of AI decisions**.
- (iii) **IEEE 7010-2020 - Well-being Assessment:** establishes quantifiable metrics to assess the impact of AI systems on human well-being. It proposes indicators to see whether an AI application is contributing to or detracting from the well-being of people affected by it, seeking to align technologies with improved quality of life.
- (iv) **IEEE 7002-2022 - Data Privacy in AI Processing:** focuses on defining privacy-preserving requirements in AI applications (PETs), including anonymization techniques, data minimization and re-identification risk assessment.

In addition to the 7000 family, the IEEE produces technical standards in basic areas for AI, such as neural networks, sensor interoperability, etc.

In 2022, the IEEE launched IEEE 2801-2022, **a standard for evaluating the document transparency of machine learning systems - more focused on *Model Cards*-type documentation.**

As with ISO/IEC, companies can voluntarily adopt these standards to guide their practices and demonstrate adoption/respect for the obligations laid down in regulations.

- **Other international standards and initiatives**

In addition to ISO/IEC and IEEE, there are consortia and industry groups defining their own guidelines for responsible AI, such as:

- a) **Partnership on AI:** formed by a consortium of companies and NGOs. It publishes a guide to best practices, such as the explainability of algorithms and the use of AI in the media and the risks associated with *deepfakes*.
- b) **W3C (World Wide Web Consortium):** discusses web standards that can influence AI, such as standardized representations for explaining automated decisions or ontologies for learning data.
- c) **OpenAI, Google, Microsoft and other companies considered PROVIDERS** release open methodologies as ***Model Cards* (model cards describing the objectives, performance and limitations of their proprietary AI models)**. This is considered a good transparency practice.

XII- International AI Governance Frameworks and Guidelines

In addition to international regulations and technical standards, AI governance can also be guided by a series of **ethical principles and frameworks developed by international bodies, governments and standardization organizations**. These guidelines are generally non-binding and serve to guide internal policies and practices, establishing a consensus on values and procedures for reliable, safe and human-centered AI.

- **OECD AI Principles**

The AI principles established by the OECD are considered a global framework. In 2019, the member countries of the OECD (Organization for Economic Cooperation and Development), including the US and many EU member states, along with other countries (totaling 42 countries), adopted five guiding principles for trustworthy AI:

- 1) **INCLUSIVE GROWTH, SUSTAINABLE DEVELOPMENT AND WELL-BEING:** AI should contribute to the inclusive advancement of society, benefiting the widest possible spectrum of people and respecting the environment, rather than deepening existing inequalities.
- 2) **HUMAN RIGHTS AND DEMOCRATIC VALUES, INCLUDING FAIRNESS AND PRIVACY:** AI systems must respect fundamental rights, the rule of law, democratic values and non-discrimination. This includes ensuring **fairness** in outcomes and equitable treatment of individuals and groups, as well as the possibility of human review of automated decisions that may affect their rights.
- 3) **TRANSPARENCY AND EXPLAINABILITY:** There must be sufficient transparency about AI systems, so that those affected can understand when they interact with an AI and get explanations about its algorithmic decisions, especially in contexts that could have an impact. This improves **contestability** and trust, as users and regulators are able to scrutinize the functioning of the systems at some level.
- 4) **ROBUSTNESS, SECURITY AND SAFETY:** AI systems must be technically robust and secure throughout their life cycle, such as testing and quality assurance to prevent intentional damage (attacks and hackers) or unintentional damage (failures or incorrect results). This includes resilience in the face of adversarial attacks, data reliability and the possibility of interrupting the system's operation if necessary - *kill switch* - as well as other security controls and measures.
- 5) **ACCOUNTABILITY (Accountability):** Mechanisms must be in place to assign responsibility and audit AI systems. AI developers and operators need to be responsible for the proper functioning of their systems in line with the principles highlighted above, and be accountable and auditable in the event of non-compliance.


These principles are so relevant that they have also inspired the **G20 AI Principles and other national and sectoral policies, such as the European Union's AI Strategy and the draft EU AI ACT itself.**

In addition to the 5 guiding principles, the OECD also sets out 5 recommendations for policymakers, including:






- **INVESTING IN AI RESEARCH AND DEVELOPMENT:** Governments should consider long-term public investment and encourage private investment in AI research and development, including interdisciplinary efforts, to stimulate innovation and promotion of responsible and trustworthy AI, with a focus on challenging technical issues and related social, legal and ethical implications. Similarly, governments should consider public investment and encourage investment in open data sets that are representative and respect privacy and data protection, to support an AI research and development environment free of inappropriate bias and to improve interoperability and the use of standards.
- **FOSTERING AN INCLUSIVE AI-ENABLING ECOSYSTEM:** Governments should promote the development of and access to an inclusive, dynamic, sustainable and interoperable digital ecosystem for reliable AI. Such an ecosystem includes, among others, data, AI technologies, computing and connectivity infrastructure, and mechanisms for AI sharing.
- **SHAPING IN ENABLING INTEROPERABLE GOVERNANCE AND POLICY ENVIRONMENT FOR AI:** Governments should foster an agile policy environment that supports the transition from the research and development phase to the deployment and operation phase of reliable AI systems. To this end, they should consider using experimentation to provide a controlled environment in which AI systems can be tested and scaled up as appropriate. Governments should review and adapt, as necessary, their policies and regulatory frameworks to encourage innovation and competition for reliable AI.
- **BUILDING HUMAN CAPACITY AND PREPARING FOR LABOUR MARKET TRANSFORMATION:** Governments must work closely with stakeholders to prepare for the transformation of the world of work and society. They must empower people to effectively use and interact with AI systems across their full range of application, including by equipping them with the necessary skills. Governments should take measures, including through social dialogue, to ensure a just transition for workers as AI is deployed, such as through lifelong training programs, support for those affected by displacement, including through social protection, and access to new opportunities in the labor market.
- **INTERNATIONAL CO-OPERATION FOR TRUSTWORTHY AI:** Governments, including developing countries and stakeholders, should actively cooperate to promote these principles and progress responsible stewardship of AI. Governments should work together with the OECD and other global and regional forums/groups to promote knowledge sharing on AI, as appropriate. They should encourage international, cross-sectoral and open initiatives, involving multiple stakeholders, to gather long-term expertise on the subject. Governments should

also encourage the development and own use of internationally comparable indicators to measure AI research, development and deployment, and gather evidence base to assess progress in implementing these principles.

Values-based principles

	Inclusive growth, sustainable development and well-being >
	Human-centred values and fairness >
	Transparency and explainability >
	Robustness, security and safety >
	Accountability >

Recommendations for policy makers

	Investing in AI R&D >
	Fostering a digital ecosystem for AI >
	Providing an enabling policy environment for AI >
	Building human capacity and preparing for labour market transition >
	International co-operation for trustworthy AI >

• Recommendation on the Ethics of Artificial Intelligence - UNESCO

UNESCO - the UN agency for education, science and culture - approved its *global standard Recommendation on the Ethics of Artificial Intelligence* in 2021, agreed by 193 member states. It is an extensive document outlining values, principles and recommended actions for countries to implement to ensure that AI is developed in accordance with human rights and universal values.

The document:

- It defines fundamental values that should permeate the entire AI, such as respect, protection and promotion of human rights, diversity and inclusion, harmony with the environment (sustainability) and guaranteeing peace;
- It establishes principles such as **proportionality and non-maleficence; security and cybersecurity; justice and non-discrimination; transparency and explainability; privacy and data protection**, among others. There are around 10 guiding principles.
- In addition to the principles, UNESCO also details **concrete policy measures that countries should pursue, such as:** ethical impact assessments before implementing AI systems on a large scale (especially in public

agencies/authorities); promoting AI education and literacy for the public; frameworks for lifecycle management and international cooperation for governance. The recommendation also suggests **creating an international registry of trustworthy AI systems, encouraging soft laws and codes of conduct.**

- It also addresses specific areas, such as ensuring that AI does not exacerbate gender disparities, such as stereotypes in virtual assistants, and that digital cultural heritage is respected.

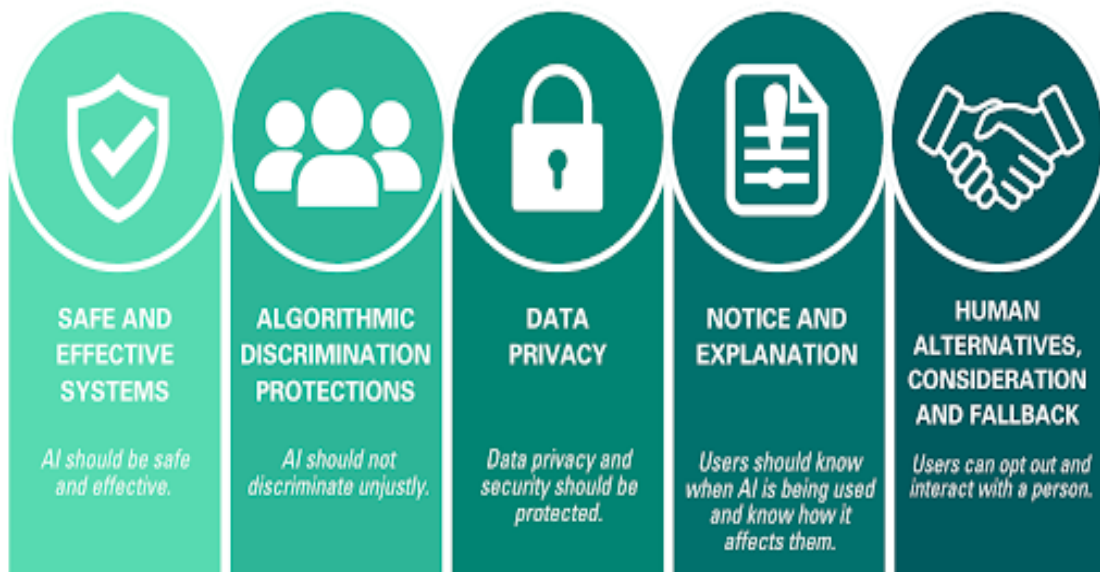
Although not binding, UNESCO's recommendations serve as a reference, especially for developing countries, to formulate their AI approaches and strategies.

- **AI Bill of Rights - White House Blueprint**

In October 2022, the OSTP - **Office of Science and Technology Policy** - issued the **Blueprint for an AI Bill of Rights**, a set of non-binding principles that aim to protect the public from potential harm from AI. This blueprint defines 5 rights or expectations that citizens should have in relation to AI systems:

- 1- **SAFE AND EFFECTIVE SYSTEMS:** The right to be protected against unsafe or faulty AI systems. Authorized systems must be tested for risks, designed to consider contexts of use and accompanied by continuous post-deployment monitoring to ensure safety throughout their use and that they do not cause unexpected and harmful results;
- 2- **ALGORITHMIC DISCRIMINATION PROTECTION:** The right not to be discriminated against by unfair algorithmic decisions on the basis of characteristics such as race, gender, ethnicity, etc. It implies that developers / PROVIDERS must audit and mitigate biases in systems and regulatory bodies must oversee discriminatory practices;
- 3- **DATA PRIVACY:** The right to protection and privacy over your personal data used in AI systems. This means the option of consent where appropriate, limits on the collection of data beyond what is necessary (*data minimization*) and incorporating privacy-enhancing techniques (PETs, anonymization, synthetic data, etc.).
- 4- **NOTICE AND EXPLANATION:** The right to know when one is interacting with an AI (and not a human being) and to receive understandable explanations of what main criteria led to an adverse automated decision. This echoes the principle of transparency: for example, if a loan applicant is turned down by an algorithm, they should be informed that the decision was automated and given a basic explanation - insufficient income (example).
- 5- **HUMAN ALTERNATIVES, CONSIDERATION AND FALLBACK:** The right to choose, whenever possible, not to be subject only to an AI system. In other words, access to a human service or review. It also means that if an automated system fails or is

not adequate, there must be contingency plans for human intervention or alternative methods of achieving the service in question.



The blueprint did not create **legal rights in the US** - it only serves as a **federal guideline and recommendation for good practice**. In other words, these rights were not legally guaranteed to the US population. Government agencies were encouraged to apply it in their spheres. For companies and developers, the blueprint is a sign of what American regulators consider to be the basis of responsible AI.

- **Human Rights, Democracy and Rule of Law Assurance Framework (HUDERAF)**

The **Human Rights, Democracy and Rule of Law Assurance** (HUDERAF) is a framework proposed by the Council of Europe to ensure that AI systems comply with human rights, democratic principles and the rule of law.

HUDERAF combines human rights due diligence processes with technical governance. Broadly **speaking, it suggests that organizations conduct AI impact assessments focused on human rights, democracy and the rule of law - AI HUMAN RIGHTS IMPACT ASSESSMENT**. This includes:

- Identify the risks that an AI system can pose to rights (such as privacy, freedom of expression, equality, etc.);
- Involve potentially affected parties in this process;
- Mitigate the risks identified;
- Document and monitor measures.

The framework also encourages companies and governments to establish **AI oversight mechanisms (multidisciplinary committees, periodic independent audits)** and to

adopt due diligence principles similar to those applied in other areas of corporate responsibility.

In 2022, the Council of Europe implemented a tool called **HUDERIA (Human Rights, Democracy and Rule of Law Impact Assessment)** aligned with HUDERAF, offering a structured guide to assessing the impacts of AI on these pillars.

- **NIST AI RMF Framework**

NIST AI Risk Management is a framework designed to help organizations manage the risks associated with AI and promote the development and use of reliable and responsible AI systems. Launched in January 2023, the framework aims to provide a flexible, non-sector-specific and use-case agnostic approach for organizations of all sizes.

Considering the transformative potential and the risks that can negatively impact individuals, organizations and society, the AI RMF emphasizes the need for proactive management to build a trustworthy AI culture.

According to AI RMF, a reliable AI (TRUSTWORTHY AI) has 7 main characteristics:

- 1- **VALID AND RELIABLE:** Guarantees that the AI is fit for purpose and the specific intended use/application has been fulfilled. In addition, the AI has the ability to function as intended, without failure, for a given time interval, under given conditions. In this sense, AI systems that prove to be inaccurate or do not generalize well to data and scenarios beyond their training, potentiate negative risks.
- 2- **SAFE:** A safe AI system is one that, under defined conditions, does not lead to a state in which human life, health, property or the environment could be threatened. Safe operation is enhanced through responsible practices *by design*. Safety risks that represent the potential for serious harm or death require the most urgent prioritization and the most complex risk management processes.
- 3- **SECURE AND RESILIENT:** The "resilience" of an AI system refers to its ability to withstand unexpected adverse events or changes in its environment or use. In addition, the system is also expected to be able to maintain the confidentiality, integrity and availability of information, through mechanisms that prevent unauthorized use. Common security concerns include adversarial attacks, data poisoning and exfiltration of models or training data.
- 4- **ACCOUNTABLE AND TRANSPARENT:** The AI system is expected to be transparent, making its results available to the individuals who interact with it. Meaningful transparency provides appropriate levels of information based on the stage of the AI lifecycle and adapted to the role/knowledge of the AI actors. Transparency ranges from decisions and actions *by design* to training data and the structure of the model, its intended use cases, etc. As for accountability, there should be clarity about who is responsible for the decisions of the AI system and that reparation of damages is allowed when negative outcomes or impacts occur.

- 5- EXPLAINABLE AND INTERPRETABLE:** Explainability refers to a representation of the mechanisms underlying the operation of AI systems. In turn, interpretability refers to the meaning of the output of AI systems in the context of their designed functional purposes. AI systems that are more "explainable" can be debugged or monitored more easily, and lend themselves to more thorough documentation, auditing and governance.
- 6- PRIVACY-ENHANCED:** AI systems must respect the norms and practices that help safeguard human autonomy, identity and dignity. Privacy values such as anonymity, confidentiality and control should guide choices *by design*. *Privacy Enhancing Technologies* (PETs) for AI, as well as data minimization methods such as de-identification and aggregation for certain model outputs, can support the design of privacy-enhanced AI systems.
- 7- FAIR:** Fairness in AI includes concerns about equality and equity, addressing issues such as prejudicial bias and discrimination. *Fairness* standards can be complex and difficult to define, as perceptions of what would be "fair" can vary between cultures and can change depending on the application. A system in which prejudicial biases are mitigated is not necessarily fair - it may still be, for example, not very accessible to individuals with disabilities or exacerbate existing disparities. There are three categories of bias that must be considered and managed:
- a. **SYSTEMIC BIAS:** Commonly present in AI datasets, organizational norms, practices and processes throughout the AI lifecycle.
 - b. **COMPUTATIONAL AND STATISTICAL BIAS:** Can be present in AI datasets and algorithmic processes, and often stems from systematic errors due to unrepresentative samples.
 - c. **HUMAN-COGNITIVE BIAS:** Relates to how an individual or group perceives and interprets the results of AI for decision-making or how the purpose and goals of an AI system are perceived.

In addition, the AI RMF also establishes 4 macro functions that organizations should adopt when assessing risks applicable to AI systems:

- **GOVERN:** permeates all other functions and levels of the organization. An AI risk management culture must be cultivated and implemented within the organization. This involves establishing policies, processes, procedures, organizational structures and responsibilities for anticipating, identifying and managing risks.
 - Understand, manage and document legal and regulatory requirements related to AI;
 - Integrate the fundamental characteristics of TRUSTWORTHY AI into the organization's policies and practices;

- Define processes to determine the necessary level of risk management activity based on the organization's risk tolerance;
 - Establish and document the risk management process and its results in a transparent manner;
 - Monitoring and periodic review of risk management processes, clearly defining roles and responsibilities;
 - Maintain an inventory of AI systems, with resources allocated according to risk priorities;
 - Provide training in AI risk management so that employees can perform their duties properly;
 - Prioritize workforce diversity and foster a mindset of critical thinking and prioritization of safety in the design, development, deployment and use of AI systems;
 - Establish processes to collect and integrate feedback from external actors (e.g. users and impacted communities) on the risks and impacts of AI.
- **MAP:** Establish the necessary context for identifying and framing AI-related risks.
- Understand and document the intended purposes of the AI system, potential beneficial uses, context-specific laws and regulations, deployment scenarios, types of users impacted and expectations;
 - Identify the negative and positive impacts of the potential use of AI systems on individuals, communities, organizations and society;
 - Document the knowledge limits of the AI system and how its output can be used and supervised by humans;
 - Identify and document the potential costs (including non-monetary) resulting from errors expected or realized by an AI system;
- **MEASURE:** Consists of adopting quantitative and qualitative techniques, methodologies and actions to analyze, evaluate, benchmark and monitor AI risks and their potential positive and negative impacts.
- Select and implement approaches and metrics for measuring the AI risks listed in the **MAP** stage, prioritizing the significant risks;
 - Properly document risks or reliability characteristics that will not or cannot be measured;

- Regularly evaluate and update the adequacy of metrics and the effectiveness of existing controls;
 - Ensure that evaluations involving human beings meet the applicable requirements and are representative of the relevant population;
 - Monitor the functionality and behavior of the AI system when in production;
 - Establish feedback with users and society who report problems and ask for automated decisions to be reviewed, integrating this feedback into evaluation and monitoring metrics.
- **MANAGE:** Involves allocating resources to address mapped and measured risks on a regular basis. It includes planning how to respond to, recover from and communicate about incidents or risk events.
- Determine whether the AI system achieves its stated aims and objectives and whether or not its development or deployment should continue;
 - Prioritize the treatment of documented AI risks based on impact, probability and available resources;
 - Develop, plan and document responses to AI risks considered to be of high priority. Likewise, document negative residual risks;
 - Plan and prepare strategies to maximize the benefits of AI and minimize the negative impacts;
 - Establish mechanisms and assign responsibilities for replacing or decommissioning AI systems that demonstrate behavior or performance inconsistent with their intended use;
 - Monitor the risks and benefits of third-party AI and apply risk controls;
 - Integrate measurable activities for continuous improvement into AI system upgrades;
 - Communicate incidents and errors to the relevant actors, including affected communities and users, as well as control and inspection authorities/agencies.



DOMAIN III

Understand how to govern AI development

I- Governing the design and construction of the AI model

1. Understanding the development lifecycle

The development of AI systems generally comprises interconnected phases of: **Planning (Plan)**, **Design (Design)**, **Development (Develop)** and **Deployment (Deploy)**.

- **PLANNING:** lays the foundations for the project. Governance is manifested in the clear definition of the business problem that AI aims to solve, alignment with the organization's mission and objectives, identification of gaps that AI can fill, preliminary assessment of data availability and suitability, definition of the project scope and establishment of the initial governance structure, including identification of a *Champion* or "sponsor" to ensure support and resources.
- **DESIGN:** Where the system's architecture begins to take shape. Governance is intensive here, focusing on implementing the data governance strategy (quality, collection, preparation, labeling, cleaning and privacy aspects), determining the system's technical architecture (choice of algorithms and platforms), conducting impact assessments and explicitly applying *by design* principles.
- **DEVELOP:** Here, the model is actually built and trained. Governance focuses on ensuring that data is collected and used in accordance with established legal standards and ethical principles, that testing is comprehensive (unit, integration, validation, performance, security, bias and interpretability), that data lineage and provenance are maintained, and that risks identified during training are monitored and mitigated. Documentation of the training and testing processes is an essential part of this.
- **DEPLOYMENT:** Involves releasing the model into production and its ongoing monitoring and maintenance. Governance includes the final assessment of readiness for release, establishing maintenance and upgrade schedules, carrying out periodic audits, managing incidents and risks and analyzing root causes of failures, and ensuring transparency through public disclosure and technical documentation.

In all phases, it is also recommended to identify the correct stakeholders for each phase - Legal, Compliance, Ethics, Privacy, Info Security, Data Science, Engineering, IT and so on.

In addition, *gate reviews* (checkpoints between phases) also help to ensure that governance requirements are met.

2. Definition of the business problem, the link with the strategic objective, the feasibility and objectives of the project

Firstly, it is vital that organizations understand precisely the problem they are trying to solve by developing the AI system. Without a thorough understanding of the problem and the associated objectives, there is a risk of developing a solution that is technically impressive but fails to deliver real value or, worse, introduces new, unanticipated risks.

From the point of view of governance, the aim at this stage is for this definition to be structured, collaborative and aligned with the organization's overall strategy.

Methodologies such as **Design Thinking** can be useful for empathizing with end users and deeply understanding their needs and pains. Similarly, user interviews, journey mapping and root cause analysis (such as the 5 whys?) help the organization to go beyond superficial symptoms and identify the fundamental problem. Here, the following questions can be asked:

- **What specific process or result are we trying to improve?**
- **Who is affected by this problem?**
- **What is the current impact of the problem?**

Likewise, it is essential that the project's objectives are directly linked to the organization's broader strategic objectives. This ensures that the project contributes to the company's overall mission and thus makes it easier to get support for its execution. With this, a **clear demonstration of the expected business value (ROI, cost reduction, revenue increase, improved customer experience) becomes essential to justify the investments, resources allocated and, consequently, the time that will be spent.**

It is also important to assess the **VIABILITY OF THE PROJECT**. This includes:

- **Technical feasibility:** Does the organization have the necessary data (in quantity and quality), technical expertise (data scientists and AI engineers) and infrastructure capable of supporting the project?
- **ECONOMIC VIABILITY:** Is the cost of developing and maintaining the AI system over the long term justified by the expected business value? A detailed cost-benefit analysis should be carried out.
- **OPERATIONAL VIABILITY:** Can the system be integrated into existing processes? Does the organization have the capacity to manage and maintain the system in the long term?
- **LEGAL AND ETHICAL VIABILITY:** Does the intended use of AI comply with applicable laws and regulations? Are there significant ethical risks that need to be addressed and eradicated before implementation?
 - **For example:** An organization intends to develop a project for the use of an AI system that is capable of evaluating behavior and manipulating likely customers in order to induce them to buy products in its e-commerce. Just by evaluating the purpose, it is understood that the project, under the

regulatory aspect of the EU AI ACT, already falls within the scope of a prohibited practice and cannot be carried out.

Another aspect that should be assessed is the definition of objectives and aligned KPIs. In this way, organizations can, for example, adopt the **S.M.A.R.T.** framework:

Specific
Measurable
Achievable
Relevant
Time-bound

In this case, instead of having the objective of "improving customer service", a SMART objective would be "reducing the average waiting time on the phone by 20% over the next 12 months by implementing an AI *chatbot* to answer frequently asked questions on our website".

Finally, a question that must be asked internally is whether to **DEVELOP vs. BUY**. Organizations have the option of either developing the AI models they deem necessary internally or acquiring them ready-made from suppliers/partners. Governance requires a careful assessment of the risks and benefits of each of these approaches. Developing the AI system in-house, despite being more complex, offers more control and monitoring over all phases, but ends up requiring resources and expertise. Acquisition by a partner/supplier tends to be faster, but introduces security, privacy and compliance risks for the supplier/partner which, depending on the internal governance of the contracting company, may become more latent. For the purchase/acquisition of AI systems by partners/suppliers, it is recommended to carry out rigorous *due diligence*, information security practices and controls and to formalize appropriate contractual clauses.

3. Types of AI solutions, applications and governance

Once the problem has been addressed and the objectives clearly defined, the next crucial stage in the *design* is the selection of the appropriate type of AI solution and its underlying technical architecture. The choice should not be based solely on technical sophistication, but rather on the suitability of the problem, the available data, resource constraints and, crucially, the governance considerations associated with each approach.

- **CLASSIFICATION SYSTEMS:** Assign items to predefined categories. Examples include email *spam* filters, simple medical diagnoses, sentiment analysis (positive/neutral/negative), simple image recognition, etc.
 - **Governance aspects:** There are risks of discriminatory bias (e.g. incorrect classification of certain demographic groups), the need for explainability (why was a particular email classified as *spam*?) and robustness to adversarial attacks (inputs manipulated to deceive the classifier).

- **REGRESSION SYSTEMS:** They predict a continuous numerical value and can be used to forecast stock prices, estimate product demand and sales, forecast the weather, credit risk.
 - **Governance aspects:** Accuracy and reliability of decisions is largely linked to the quality of the training data. In addition, there are questions about the explainability of the factors influencing the decision, the stability of outliers in the data and the potential use for decisions with a high impact on people, such as credit analysis and insurance.
- **RECOMMENDATION SYSTEMS:** Suggest relevant items to users, based on previous use. Examples include recommending movies (Netflix), products (Amazon and Mercado Libre), music (Spotify) and social networks (personalized feed).
 - **Governance aspects:** There is a risk of creating **echo chambers**, a phenomenon present in recommendation algorithms that learn to display predominantly content that confirms the user's pre-existing preferences or beliefs, reducing the diversity of points of view and potentially increasing radicalization and social polarization on sensitive topics. The system "echoes" the same messages. This risk is accentuated in recommendation systems, as many of them are designed to optimize engagement, using display time and click metrics to generate "emotional" responses. In addition to *echo chambers*, there are also risks applicable to the transparency of how recommendations were generated, potential manipulation of users (growing risk related to THEORY OF MIND) and privacy risks related to profiling/segmentation of holders (including behavioral analysis).
- **GENERATIVE SYSTEMS:** Create new content, such as text, images, audio and code. Examples include the most popular LLMs, such as ChatGPT, Gemini, Claude, DALL-E, Midjourney and ElevenLabs.
 - **Governance aspects:** Risks of copyright infringement (model training using protected data, generation of derivative content), disinformation and *deepfakes*, toxic or biased content, plagiarism, security and privacy, limited and complex explainability (black-box) and intensive use of computing resources (energy and sustainability).

In addition to the choice of system and its respective governance-related risks, the specific architecture must also be chosen, which could be CNN, RNN, Transformer, etc.

In the context of AI model development and governance, architecture **is the physical-logical design of the model itself - in other words, how the neurons, layers and internal connections are organized to transform input into output.**

Specifically for neural networks, the main types of architectures are:

- **Feed-Forward (FNN / MLP):** Data travels through dense layers in a single direction, without cycles.

In other words, the data enters through one door, passes through a row of tables and leaves through another door. There are no cycles or other paths.

Generally used to predict credit risk, classify emails as *spam* or not or any data in a table.

- **Convolutional (CNN):** Layers apply local filters that are repeated over the entire input.

Imagine a team of researchers from antiquity using magnifying glasses - each one analyzes small pieces of an image and highlights patterns (edges, colors, shapes). Then each of them puts it all together and makes a final assessment.

Generally used for photo recognition (e.g. detecting cracks or potholes in roads), automatic reading of x-rays or other images for medical purposes, quality control by camera in a factory.

- **Recurrent (RNN / LSTM / GRU):** It not only processes current information, but also takes into account what it has learned from previous information in the sequence. Information can "circulate" within the network, allowing it to remember past events when processing new ones. In other words, it has a "memory" to facilitate current data processing.

It is often used in automatic text translation systems, sentiment analysis on social networks and daily energy demand forecasting.

- **Graph Neural Network (GNN):** model entities as *nodes* and their relationships as *edges*, letting information propagate across the graph. It doesn't just look at single items, but at how they are connected in a kind of network or "graph". It analyzes each "node" and "edge" - the information from one "node" can influence its neighbors, causing the GNN to learn from this relationship structure.

It is used, for example, in complex relationship systems, such as social networks, where it suggests a friend or a person you probably know, based on your friendship with other people in a close social circle, or in recommendation systems, where it suggests products based on what people with similar tastes to you have bought.

These are technical points that require evaluation by data scientists and AI engineers, but AI governance can also influence the choice of architecture, taking into account the requirements of explainability (simpler models, such as decision trees, are preferable in some regulatory contexts), robustness and computational efficiency.

The decision on the chosen architecture must therefore be taken into account:

- **TECHNICAL PERFORMANCE** - taking into account precision, recall, robustness and stability of results;
- **INTERPRETABILITY/EXPLICABILITY** - how easy it is to understand how the model makes decisions;
- **DATA REQUIREMENTS** - amount and type of data required for training;
- **SCALABILITY** - the ability to handle growing volumes of data and users/owners;
- **RESTRICTIONS** - development time, computational cost and available technical expertise;
- **COMPLIANCE** - Compliance with legal/regulatory requirements and obligations;
- **ETHICAL RISKS** - Potential bias, negative social impacts.

In conclusion, as a governance aspect, it is recommended that the decision on the type of system and architecture chosen be documented and justified, taking into account all the relevant dimensions, not just raw technical performance.

4. Conducting comprehensive impact assessments

A central pillar in the governance of AI in the design and development phase is the proactive conduct of impact assessments. The aim of these assessments is to understand and mitigate the potential negative effects of the AI system on individuals, groups, organizations and society as a whole, covering ethical, legal and social dimensions.

- **Algorithmic Impact Assessment (AIA):** Focuses specifically on the risks introduced by the use of algorithms, particularly in terms of *fairness*, bias and discrimination.
- **Data Protection Impact Assessment (DPIA):** Required by the GDPR and other regulations of the same nature, the DPIA focuses its analysis on the risks applicable to the processing of personal data in high-risk activities.
- **Human Rights, Democracy and the Rule of Law Impact Assessment (HUDERIA):** Proposed by the Council of Europe and aligned with HUDERAF - Human Rights, Democracy and The Rule of Law Assurance Framework - this assessment seeks to evaluate the potential impacts of AI systems in these fundamental areas. There are also tools such as the **RIN - Risk Index Number** and the Mitigation Hierarchy.
- **Fundamental Rights Impact Assessment (FRIA):** Similar to HUDERIA, FRIA focuses on fundamental rights enshrined in international treaties and frameworks.
- **Ethical Impact Assessment (EIA):** Evaluates the system's alignment with ethical principles (justice, autonomy, non-maleficence, beneficence, explainability, etc.), identifying potential ethical dilemmas.

In general, conducting an effective impact assessment generally follows a process that can be structured as follows:

- **Defining the scope:** identifying which AI system is being evaluated, its purpose and context of use;
- **Identification of risks and impacts:** Brainstorming and systematic analysis of potential harm in different dimensions (privacy, justice, security, autonomy and human rights);
- **Risk analysis:** Assess the probability and severity of the risks identified (using methodologies such as HUDERIA's risk matrix or RIN);
- **Identification of mitigation measures:** Define technical, procedural, security and organizational controls to prevent/mitigate AI system risks;
- **Consultation with relevant stakeholders:** Involve internal and external experts, as well as representatives of potentially affected groups, to obtain feedback and understand other perceptions;
- **Documentation:** Record the entire *design/development* process, including risks, analysis, controls and mitigation aspects, as well as decision-making;
- **Review and approval:** Submit the document for evaluation by a committee or to a responsible body for review and approval, always assessing the risk of conflict of interest;
- **Monitoring and reassessment:** Impact assessment is not an action with a single end point. It should be revisited periodically and/or whenever there is a significant change in the system or its context of use.

In addition, with regard to the structure of these risk assessments, organizations can use ready-made and publicly accessible *templates* provided by either supervisory authorities/agencies (such as the iconic DPIA *template* provided by the ICO) and/or use/adapt technical risk management standards (such as ISO/IEC and IEEE), as well as established frameworks such as the NIST AI RISK MANAGEMENT FRAMEWORK. Many technology companies also provide such documentation publicly, such as Microsoft and Google.

In any case, it is important to note that, even with the ease of use of the accessible *template*, it is recommended that organizations adapt their impact assessments to the existing regulatory requirements in each region, the organization's own risk strategy, sector of operation and other contexts that may vary.

5. Applying risk assessment methodologies

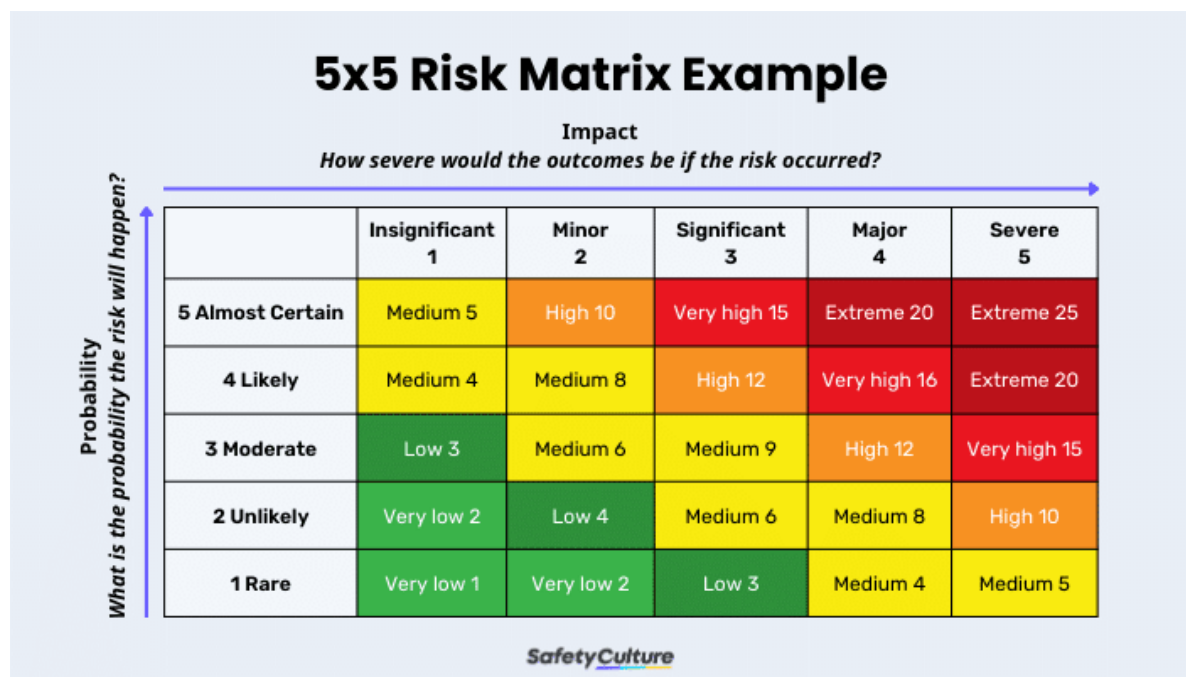
After identifying the potential negative impacts during the impact assessment, the next step in AI system *design* governance is systematic risk assessment. This involves analyzing the **probability of occurrence and the severity of occurrence, allowing the most critical**

risks to be prioritized and priority mitigation resources to be allocated more effectively.

Various methodologies can be adopted here:

- **Probability and severity matrix (e.g. 3x3 or 5x5):** visual and widely used tool. It classifies risks in two dimensions: **PROBABILITY** (low, medium, high and very high) and **SEVERITY** (insignificant, marginal, moderate, critical or catastrophic).

The combination of these two factors positions the risk in the matrix, which generally uses colors to identify risk levels.



- **HUDERIA's Risk Index Number (RIN):** this analysis, included in HUDERIA, offers a granular approach focused specifically on the risks applicable to human rights. It consists of:
 - **GRAVITY POTENTIAL:** assess the maximum severity of the potential harm to human rights, dignity or integrity (scale 1-4: Moderate / Minor, Serious, Critical and Catastrophic);
 - **RIGHTS-HOLDERS AFFECTED:** quantifies the number of people potentially affected (scale 0.5 to 2.0, based on number ranges such as 1-10k, 10k-100k etc).

With these two factors, it is possible to arrive at:

- **SEVERITY:** calculated as the sum of **GRAVITY POTENTIAL** and **RIGHT-HOLDERS AFFECTED**;
- **LIKELIHOOD (Probability):** assesses the chance of damage occurring (scale 0-4: not applicable; unlikely; possible; probable; very probable);

And with that, we arrive at RIN:

- **RIN (RISK INDEX NUMBER):** Result of the sum of **SEVERITY** and **LIKELIHOOD**, determining the risk level as **LOW (less than 5)**; **MODERATE (5.5 to 6)**; **HIGH (6.5 to 7.5)** and **VERY HIGH (greater than 8)**.

Finally, after identifying the RIN and the applicable risks, the organization must conduct a **Mitigation Plan, which uses a hierarchy to choose the measures in proportion to the severity/RIN**. To do this, 4 GRADUAL levels are defined (the first being the ideal and the last the least recommended):

- (i) **AVOID:** Modify the design or development to completely eliminate the risk. Here, you want to prevent the impact before it exists;
- (ii) **REDUCE:** Implement technical, organizational or procedural safeguards (e.g. input filters, usage limits and human control) to minimize the magnitude or probability of the impact;
- (iii) **RESTORE:** After one year, to return people to the same or equivalent situation as before (EXAMPLE: public retraction, correction of registration, restoration of a denied right).
- (iv) **COMPENSATE: Compensate**, in value, when avoidance/reduction/restoration is not enough or not possible.

6. Incorporating ethical design principles

In addition to identifying and mitigating risks and analyzing legal/regulatory compliance when building AI systems, it is also important to proactively incorporate *ethics by design* into the design and development process itself. These principles serve as a moral compass, guiding teams in making decisions that promote fair, safe and beneficial results for society as a whole.

Some of the most important principles:

- **FAIRNESS:** Ensuring that the AI System does not produce results that are systematically biased or discriminatory against individuals or groups, especially those belonging to protected categories. This involves the careful analysis of training data, the selection of appropriate *fairness* metrics (EX: demographic parity,

equal opportunity, equalized accuracy) and the application of bias mitigation techniques.

- **SAFETY AND ROBUSTNESS/RELIABILITY:** Designing systems that operate reliably and safely within their defined operating limits, and that are resilient to failures, errors and adversarial attacks. This includes the adoption of rigorous testing, validation, continuous monitoring and fallback or *kill switch* mechanisms.
- **PRIVACY AND SECURITY:** Incorporate data protection and information security from the *design* and first construction actions of the AI system (Privacy by Design). This involves data minimization techniques, anonymization/pseudonymization, encryption, PETs (Privacy Enhancing Technologies), strict access control and continuous security testing.
- **TRANSPARENCY AND EXPLAINABILITY:** Making the functioning of the AI system understandable to different audiences. Transparency refers to clarity about how the system was developed, trained and how it operates. Explainability refers to the ability to provide understandable reasons for decisions or predictions made by the model. With this in mind, tools such as *model cards*, *data sheets* and XAI (Explainable AI) techniques such as LIME (Local Interpretable Model-Agnostic Explanations) and SHAP (Shapley Additive Explanations) are seen as essential.
- **ACCOUNTABILITY:** Establish clarity about who is responsible for the development, implementation, operation and results of the AI system. This involves defining roles and responsibilities, detailed records/documentation, establishing auditing mechanisms, ensuring that there are avenues for reviewing and challenging results/predictions and repairing damage.
- **HUMAN-CENTRIC AND OVERSIGHT:** Ensure that AI systems serve human interests and that there are appropriate mechanisms for human supervision and intervention. The choice of supervision level (human-in-the-loop; human-on-the-loop; human-out-of-the-loop) should be based on the concrete risk and the context of the application. The higher the risk of the AI system, the more active/close human involvement is recommended.
- **NON-MALEFICENCE AND BENEFICENCE:** the fundamental principle of "doing no harm" and, ideally, promoting individual and social well-being.

For these principles to be applied in practice, organizations must:

- (i) **Implement clear policies and guidelines:** Internal documents that define the organization's commitment to ethical AI and provide specific guidance for development teams.
- (ii) **Training and awareness-raising:** Educating teams about ethical principles and how to apply them in their daily work.

- (iii) **Tools and processes:** Incorporate ethical *checklists*, impact assessments and *fairness* metrics into development and review processes.
- (iv) **Ethics/Review Committees:** Adoption of multidisciplinary bodies to review high-risk AI projects and provide guidance on ethical dilemmas.
- (v) **Stakeholder engagement:** Consult areas and specialists in ethics, legal, compliance and also affected groups.

In addition, it is important for the AI Governance team to bear in mind that applying ethical principles can involve *trade-offs*. For example, increasing explainability may, in some cases, slightly reduce the accuracy of the model. Maximizing *fairness* for multiple groups simultaneously may be mathematically impossible in certain situations. Governance involves navigating these *trade-offs* in a transparent and justified way, documenting decision-making and assessing the context, risks, legal/regulatory analysis and the company's strategy on the subject.

7. Mapping and engaging stakeholders effectively

The successful development and implementation of AI systems rarely takes place in isolation and without collaboration. They depend on collaboration and alignment between a wide range of internal and external stakeholders, each with their own perspectives, knowledge, interests and concerns. Effective AI governance therefore requires a structured process for identifying, analyzing, mapping and engaging these stakeholders in an ongoing and meaningful way.

Key steps include:

- 1) **Identification and analysis:** the first step is to identify who the relevant stakeholders are.

Generally speaking, stakeholders can be **INTERNAL** and **EXTERNAL**.

INTERNAL stakeholders can be executive leadership (senior managers and the board), product teams, Legal, Compliance, Privacy, Information Security, HR, Marketing, Sales, Operations, Customer Support, Communications and other areas of the organization.

EXTERNAL stakeholders can be the customers/end users themselves, regulators, business partners and investors, members of academia, NGOs, groups and associations and the media.

After the first identification, you should analyze the relationship of each of these stakeholders in the project (whether they support, oppose or act neutrally) and the risks and opportunities they represent.

- 2) **Categorization into power and interest:** Stakeholders must be categorized in order to better adapt them to the governance process of the AI being developed. In this case, we can separate them into:
- a. **Stakeholders with HIGH power and HIGH interest:** they should be **closely involved**, consulted regularly and included in key decisions;
 - b. **Stakeholders with HIGH power and LOW interest:** must be informed about relevant decisions and avoid information overload;
 - c. **Stakeholders with LOW power and HIGH interest:** regular communication, consultation on specific areas of interest and feedback channels;
 - d. **Stakeholders with LOW power and LOW interest:** minimum communication required.
- 3) **Engagement strategy:** Engaging these stakeholders, especially those most relevant to the project, is vital for AI governance. Therefore, engagement should not only be informative, but genuinely consultative and collaborative, and should include them in the activities, deliverables and next steps.

Stakeholder engagement strategies include:

- **Customized communication:** adapting the message and channel to the profile of each stakeholder;
- **Collaborative workshops:** hold structured meetings to co-create solutions, identify risks and/or define requirements (e.g. Design Thinking sessions, participatory impact assessment sessions);
- **Multidisciplinary committees:** Establish governance or ethics committees with representation and participation from different areas;
- **Formal consultations and feedback:** structured processes for collecting feedback on specific proposals or documents, as well as concerns or questions;

With the identification and engagement of stakeholders, in conducting the AI project and governance for the system under development, it is natural for **conflicts to arise**, which can occur due to divergent priorities (e.g. speed of development vs. ethical rigor and privacy risks). Governance must establish mechanisms to identify, analyze and manage these conflicts constructively, seeking solutions that balance the different interests or escalating the decision to a higher level when necessary.

8. Evaluating use cases rigorously

It is essential that governance carries out a rigorous assessment of the use case proposed for the *design* and development of the AI system. This process goes beyond simply identifying an opportunity - it involves a critical analysis of the feasibility, potential value and strategic alignment of the AI application, ensuring that the projects selected are not only technically possible, but also strategically relevant and manageable in terms of risk.

Thus, adopting an **evaluation framework** can help ensure that all relevant aspects are considered. As a minimum, this framework should cover:

- (i) **IMPACT:** What is the potential positive impact of the AI solution on the identified business problem? Will it solve the problem completely or only partially? What is the expected quantifiable value (cost reduction, revenue increase, efficiency improvement, increase in customer satisfaction); Is the impact strategic for the organization?
- (ii) **EFFORT:** What resources are needed to implement and maintain the solution? This includes financial costs (development, infrastructure, capacity), time (projected schedule), technical expertise and data.
- (iii) **FITNESS:** How does the proposed AI solution fit into the organizational context and with the company's strategic objectives? How is the AI aligned with the organization's culture and values? Is it compatible with the organization's existing logical infrastructure and other systems/processes? Does the organization have the necessary maturity and readiness to adopt and manage this technology? What are the associated risks (technical, ethical, legal and operational) and does the organization have the capacity to mitigate them?

In addition, with the AI *hype* and the various ideas and solutions that aim to solve various internal problems, it is common for organizations' AI Governance teams to receive several queries for case ideas. As a result, **a prioritization process must be carried out to identify the projects that are most essential and feasible**. This can be done using the following methods:

- **SCORECARDS:** Assign scores to each use case based on predefined criteria (e.g. IMPACT, EFFORT, FITNESS and RISK) and rank the use cases based on the total score;
- **MoSCow (Must Have, Should have, Could have, Won't have):** Classifying use cases based on their strategic importance and urgency, always varying from organization to organization.
- **RICE (Reach, Impact, Confidence, Effort):** A quantitative method that evaluates the reach (how many users/customers affected), the impact per user, the confidence in the estimate and the effort required for implementation.

No less important, this process of analyzing and evaluating the intended use case must also include the **participation of relevant stakeholders**, who will bring insights,

recommendations and definition of associated risks, such as business, technology, finance, legal, compliance and risks.

Another relevant action can be to carry out **case studies and benchmarking with other organizations in the same sector or with similar challenges, to understand how they have implemented and evaluated similar (or identical) use cases to provide valuable insights.**

9. Comprehensive documentation as evidence of compliance

Rigorous and comprehensive documentation is the backbone of AI governance, especially during the *design* and development phase of an AI system. It serves multiple crucial purposes:

- Provides evidence of compliance with legal, regulatory and ethical requirements;
- It facilitates transparency and explainability;
- It enables and facilitates efficient auditing and accountability;
- Supports system maintenance and future updates;
- Promotes knowledge sharing within the organization;

The absence of adequate documentation can expose the organization to significant risks and undermine confidence in the system, even before its final implementation.

But what should be documented? **Documentation at this stage should cover all key decisions and processes, including:**

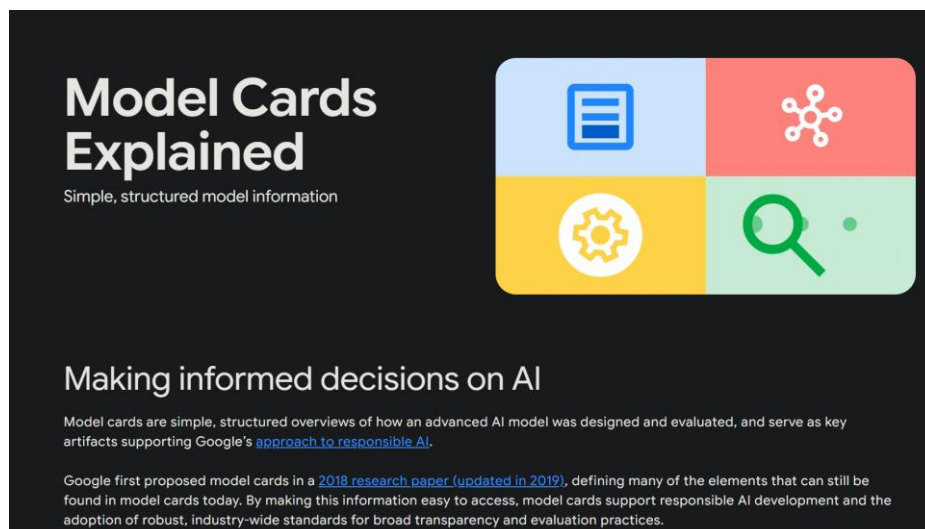
- (i) **DEFINITION OF THE PROBLEM AND OBJECTIVES:** The business case, the SMART objectives, the defined KPIs;
- (ii) **IMPACT ASSESSMENTS:** The results of impact assessments (EIA, FRIA, HUDERIA, DPIA, etc.) must be documented;
- (iii) **LEGAL AND ETHICAL REQUIREMENTS:** The applicable laws and regulations identified, the ethical principles adopted and how they have been incorporated into the design, the security and robustness measures of the model, the legal assumptions for data processing, etc;
- (iv) **STAKEHOLDER PARTICIPATION:** The participation and decision-making of the relevant stakeholders in the construction and *design* of the AI system must be documented, for control and accountability purposes;
- (v) **DATA:** Sources of **data**, legal rights regarding its use, collection, preparation and cleaning processes, quality analysis and possible biases, description of training sets, validation and testing;
- (vi) **DESIGN AND ARCHITECTURE:** Justification for choosing the type of solution and architecture, detailed description and design of the architecture, algorithms used, model parameters, keys and weights;

- (vii) **USE CASE EVALUATION:** The analysis of feasibility, impact, effort and suitability;
- (viii) **PILOT TESTS:** Plan, results and decisions based on the pilot carried out;
- (ix) **KEY DESIGN DECISIONS:** Justifications for important choices made during the process.

Still in the context of documentation, some standardized structures and tools may be advisable:

- **Model Cards:** concise documents that summarize the essential information about an AI model, including its purpose, performance, limitations, data used and ethical considerations;

- **EXAMPLES:**



<https://modelcards.withgoogle.com/>



https://github.com/huggingface/huggingface_hub/blob/main/src/huggingface_hub/templates/modelcard_template.md

- **Data Sheets for Datasets:** Proposed by Google researchers, it focuses on documenting, in detail, the datasets used to train and evaluate the models, covering maintenance, composition, collection process, pre-processing, distribution, data and potential biases. It is essential to transparency regarding the use of the data, the legality of the data and the assessment of potential biases;
- **System Cards:** Used by organizations such as OpenAI, they provide a broad view of the AI system as a whole, including not only the model, but also the infrastructure, interfaces, risks and usage policies, addressing safety and limitations in a more systemic context.
 - **EXAMPLES:**
 - [GPT-4o System Card](#)
 - [GPT-o3 and o4 mini System Card](#)
- **Counterfactual Explanations (CFEs):** Commonly related to explainability, this documentation explains why a particular prediction was made, showing what the smallest change in the input would alter the prediction. It is especially used in automated decision systems with significant impacts - credit, recruitment, medical screening, etc.

Finally, it should be noted that this documentation should be **stored in an accessible and organized place, also following the recommendations:**

- **Versioning:** adopting versions in the documentation, to keep track of changes in the documents and by whom they were made;
- **Standardization:** use consistent templates and formats;
- **Review and approval:** Implement formal review and approval processes for key documents;
- **Maintenance:** Keeping the documentation up to date as the system evolves.

10. Navigating regulatory considerations in *Design*

- **EU AI ACT:** For developers (PROVIDERS) of high-risk systems (covering areas such as biometric identification, critical infrastructure, education, employment, essential services, law enforcement, migration and administration of justice), the standard imposes strict regulatory requirements that must be considered from the *design* of the AI system:

- **Risk management system:** implement and maintain a continuous process of risk identification, analysis and mitigation throughout the life cycle;
- **Data governance:** ensuring that training, validation and testing data is relevant, representative, error-free and complete, with special attention to mitigating bias;
- **Technical documentation:** maintain detailed documentation on the system, its purpose, how it was developed and tested, and how it meets regulatory requirements. This document can even be made available not only to the supervisory authority (if requested), but also to the Deployer, who will apply the system in practice.
- **Log keeping:** Ensure that the system automatically generates logs to track its operation and enable it to be monitored and audited;
- **Transparency and information for users:** Provide clear and adequate information about the system's capabilities and limitations;
- **Human supervision:** Implement appropriate measures to enable effective human supervision;
- **Accuracy, robustness and cybersecurity:** Design and test the system to ensure adequate levels in these 3 aspects.

II- The centrality of data in AI governance

1. Assessing and documenting legal rights to use data

The first step in data governance for AI is to ensure that the organization has the legal and unambiguous right to collect, process and use the data for the intended training and testing purposes. Failure to establish a solid legal foundation can result in violations of privacy, intellectual property/copyright, contractual disputes and severe regulatory sanctions. This whole process must be well documented.

Under the Privacy and Data Protection aspects of the use of personal data, depending on the specific jurisdiction, there may be different legal bases. Under the GDPR, for example, the most common legal bases include:

- (i) **Consent**
- (ii) **Contract execution**
- (iii) **Legal obligation**
- (iv) **Public authorities**
- (v) **Legitimate Interest** - often invoked to use data for AI training.

Also, with regard to the use of personal data, it is essential to demonstrate transparency to users regarding the use of their information. To this end, it is necessary to include specific passages in the Privacy Policies and Terms of Use (TCUs) that indicate in clear language that your personal data can be used, for example, to train AI models.

Still in the context of personal data processing, the use of sensitive personal data generally requires stricter legal bases and stricter controls, considering the greater impact.

Now, under the Intellectual Property sphere, the use of data for AI training raises complex and still uncertain issues, especially related to Copyright, considering the use of protected content in generative models without permission, potentially posing significant litigation risks, similar to what happened between The New York Times v. Open AI and The Wall Street Journal v. Perplexity. In addition, there are also risks related to trade secrets, with the use of proprietary data that may contain confidential and restricted information about organizations.

2. Ensuring data quality, integrity and fitness for purpose

The maxim "*garbage in, garbage out*" highlights the critical importance of data quality in the development of AI systems. Models trained with inaccurate, incomplete, inconsistent, outdated, irrelevant and poorly represented data have produced unreliable results, potentially leading the system to make wrong, biased decisions and, consequently, harm users.

Data governance for AI must therefore **implement rigorous processes to guarantee and maintain high data quality throughout the system's lifecycle.**

- **Dimensions of data quality:** Data quality is a multidimensional concept. The dimensions include:
 - **Accuracy:** The degree to which data correctly reflects the real-world objects or events it describes. Errors can arise from incorrect measurements, *typos* or outdated data.
 - **Completeness:** Missing values in important attributes can bias analysis or hinder effective model training, consequently generating less accurate, equitable and fit-for-purpose results.
 - **Consistency:** The absence of contradictions in the data, both within the same set and in different systems.
 - **Timeliness / Currency:** The degree to which data is up-to-date and available at the time it is needed. Outdated data can lead to incorrect forecasts.
 - **Validity:** The degree to which data complies with business rules or in defined formats (e.g. an e-mail address has the correct format).
 - **Uniqueness:** The absence of duplicate records for the same entity.
 - **Relevance:** The degree to which the data is appropriate and useful for the specific purpose of the model. Using irrelevant data sets can introduce noise and degrade performance.
 - **Representativeness:** The degree to which the training data adequately reflects the diversity and characteristics of the population or environment where the model will be deployed. Lack of representativeness is a common cause of algorithmic bias.

Quality assurance processes:

- **Data Profiling:** initial analysis of the data to understand its structure, content, quality and identify potential problems (e.g. distribution of values, null values, outliers);
- **Definition of data quality rules:** Establish business rules and specific metrics for each quality dimension relevant to the use case;
- **Data Cleansing:** Iterative process to detect, correct or remove errors, inconsistencies and duplicates;

- **Data validation:** Continuous checking of data against the defined quality rules, both on ingestion and during processing;
- **Quality monitoring:** Continuous monitoring of data quality metrics over time to detect degradation.
- **Responsibility for governance:** Responsibility for data quality must be clearly defined. This can be either via *Data Stewards* or *Data Owners*. Policies and procedures for quality management must be established and communicated. Quality metrics should be reported regularly.

3. Mastering the data preparation process - *Data Wrangling*

Data preparation, often called *Data Wrangling* or *Data Munging*, is the process of transforming raw and often disorganized data into a clean, structured format suitable for analysis and training AI models. It is notoriously one of the most time-consuming and labor-intensive stages, consuming many final hours of the project.

Effective governance of this stage is crucial to guaranteeing the quality, traceability and compliance of the process.

- **Why is preparation necessary?** Raw data is rarely ready for use. It can come from a variety of sources, in a variety of formats (structured, semi-structured and unstructured), contain errors, missing values, inconsistencies and inaccuracies, or simply not be in the format required by the algorithms.
- **What are the typical steps in *Data Wrangling*?**
 - **Discovery & Exploration:** Understanding the available data, its structure, content, sources and initial limitations. This involves data profiling, descriptive statistical analysis and visualization.
 - **Structuring:** Organizing the data into a tabular or other format suitable for analysis (EX: converting unstructured logs into rows and columns);
 - **Cleaning:** Dealing with quality problems identified in the scan, such as missing values, outliers, errors, duplicates and inconsistencies;
 - **Enriching:** Combining data from multiple sources or adding new information to increase the value of the dataset (EX: adding demographic data to transaction data);
 - **Transforming:** Modifying the data to suit the model's requirements. From a technical point of view, this can include **Encoding, Feature Engineering and Aggregation**.

- **Validating:** Checking that the transformed data meets the quality and format requirements, and that the transformations have been applied correctly;
 - **Publishing/Storing:** Saving the prepared data in a format and location suitable for use by the modeling teams.
- **Data Wrangling Governance:**
 - **Documentation:** record all the preparation stages, transformations applied, business rules used and justifications for the decisions taken.
 - **Traceability / Lineage:** Maintain the ability to trace prepared data back to its original sources;
 - **Versioning:** Control the versions of the preparation scripts and the resulting data sets;
 - **Security and Privacy:** Ensure that personal data is handled appropriately during preparation and prioritize, if possible, anonymization or pseudonymization.
 - **Collaboration:** Facilitate collaboration between data engineers, data scientists and domain experts.

4. Navigating the "5Vs" of data in AI practice

The concept of the "5Vs" (Volume, Velocity, Variety, Veracity and Value) is often used to describe the characteristics and challenges in Big Data. In the context of AI development, understanding and managing each of these dimensions can also be key to effective data governance.

- **VOLUME:** Refers to the enormous amount of data generated and collected. For AI, large volumes of data are often needed to train complex models, especially *deep learning*.
 - **Governance challenges:** Storage and processing costs, the need for scalable infrastructure (cloud, distributed clusters), large-scale quality assurance, privacy risks associated with large data sets and algorithm efficiency.
- **Speed:** Describes how quickly data is generated and needs to be processed. Many AI applications require real-time or near-real-time processing.

- **Governance challenges:** Need for streaming processing architectures (Kafka, Flink, Spark), guarantee of data quality and consistency in real time, continuous monitoring of performance and drift.
- **Variety:** This refers to the different types of data that need to be managed, including structured data (relational databases, spreadsheets), semi-structured data (JSON, XML, Logs) and unstructured data (text, images, audio and video).
 - **Governance challenges:** Complexity in integrating different sources, the need for different tools and techniques for each type of data, ensuring consistency between formats, extracting features from unstructured data.
- **Veracity:** This concerns the quality, reliability and accuracy of the data. Data can be uncertain, imprecise, ambiguous or contain biases.
 - **Governance challenges:** Difficulty in assessing the reliability of external sources (web, social networks), propagation of errors, impact of low-quality data on the performance and *fairness* of the model, need for cross-checking.
- **Value:** Refers to the usefulness and business value that can be extracted from the data. Not all the data collected is valuable or relevant to a given problem.
 - **Governance challenges:** Cost of collecting, storing and processing low-value data, difficulty in identifying which data is truly predictive, need to align data collection with business objectives and the purpose of the model.

All five of these dimensions are interconnected and often involve *trade-offs*. For example, to increase processing speed, veracity can be compromised if quality checks are simplified. Dealing with high variety can increase the complexity and cost involved. **Effective governance requires a holistic understanding of these dimensions and making conscious decisions about how to balance them according to the requirements of the specific use case.**

EXAMPLE: An AI system for monitoring urban traffic. **VOLUME:** data from sensors, cameras, GPS from millions of vehicles. **SPEED:** Data arriving in real time. **VARIETY:** Numerical data from sensors, images from cameras, route data from traffic apps, weather data. **VERACITY:** Accuracy of sensors, quality of images in different conditions, reliability of user reports. **VALUE:** Ability to predict congestion, optimize traffic lights and inform drivers of routes and events/alerts.

5. Implementing data cleansing - techniques and governance

Data cleansing (or *data scrubbing*) is a relevant subset of **data wrangling**, specifically focused on identifying and correcting (or removing) errors, inconsistencies and inaccuracies in data sets. It is an iterative and often complex process, requiring technical knowledge and understanding of the problem domain. Inadequate cleaning of AI system datasets can perpetuate or even amplify quality problems, leading to unreliable models.

- **Common types of "dirt" in data:**

- **Missing values:** missing data in certain fields of the data set. Missing values break the traceability of decisions and can violate rules of fairness if the absence is concentrated in minority groups.
- **Outliers:** Extreme values that deviate significantly from the rest of the data set. An example would be a database with customer demographic information and there is a finding of people aged 200.
- **Typing and formatting errors:** Data entered incorrectly in the set or that does not follow the expected pattern.
- **Duplicates:** Identical or almost identical records for the same entity/information.
- **Inconsistencies:** Logical contradictions or violations of business rules.
- **Irrelevant data/noise:** Information that does not contribute to the analysis or modeling.
- **Toxic data:** Harmful, offensive, illegal or biased content (particularly relevant in text data or web images).

In addition, where applicable, when the data set contains personal data, anonymization and pseudonymization techniques are always recommended, for example:

- **K-anonymity:** ensuring that each record is indistinguishable from at least k-1 others);
- **L-diversity:** ensuring diversity of sensitive values within equivalent groups;
- **T-closeness:** ensuring that the distribution of sensitive values in a group is close to the general distribution.
- **Differential Privacy:** adding controlled noise to protect individual privacy.

In addition, as previously emphasized, the document process also becomes relevant, including for the data cleaning process. The organization should record which cleansing

techniques were applied, why, and what parameters were used. The impact of the cleansing and the definition of clear policies on how to deal with different types of errors should also be documented. Involving domain experts to validate cleaning decisions (considering the technical complexity of the task), especially in ambiguous cases, is also a valid action. Finally, the organization should keep logs of the changes made to the dataset and ensure that the cleansing process can be repeated consistently.

6. Establishing data lineage and provenance

Considering that data flows from multiple sources, undergoes various transformations and is used to train and evaluate different versions of models, **the ability to trace the origin and history of this data becomes fundamental**. Data **lineage** and **provenance** are key concepts for this traceability, essential for governance, auditing, debugging and reproducibility.

With this in mind, we highlight the following concepts:

- **PROVENANCE:** Refers to the origin and historical record of the data. It answers questions such as - *Where did this data come from? Who created or modified it? When? What processes were involved in its creation?*
- **Lineage:** Describes the complete life cycle of the data, mapping its flow through different systems, processes and transformations over time. It answers questions such as - *What sources fed this dataset? What transformations have been applied to this dataset? What models or applications use this dataset?*

From a governance perspective, the lineage of the data is essential:

- (i) **Auditability and compliance:** makes it possible to demonstrate the origin of the data and the transformations applied, essential for facilitating compliance with regulatory requirements and for internal/external audits.
- (ii) **Reproducibility:** Makes it possible to recreate data sets and model results - essential for scientific validation and debugging.
- (iii) **Impact analysis:** Helps to understand how changes in data sources or transformation processes can affect *downstream* models or reports.
- (iv) **Debugging:** Facilitates the identification of the root cause of errors or biases in the model, tracing them back to the original problematic data or processes.
- (v) **Quality management:** Identifies low-quality data sources or processes that have introduced errors into the model.
- (vi) **Trust and transparency:** Increases confidence in data and models by providing visibility into their history.

7. Conducting comprehensive testing strategies

Testing models goes far beyond measuring accuracy or performance. Given the complexity and potential impact of AI systems, a comprehensive, multi-faceted testing strategy is also essential to ensure robustness, security, *fairness* and compliance with requirements and purposes. Governance requires that testing is planned, executed and documented on a regular basis.

- **Factors that guide testing:** the testing strategy should be informed by:
 - **Identified risks:** focus tests on the areas of greatest risk identified in the impact assessments;
 - **System purposes:** more rigorous testing is required for high-risk systems or those with the potential to have a critical impact on users;
 - **Type of algorithm:** Different types/modalities of algorithms may require specific types of tests;
 - **Integration with third parties:** Testing the interaction with external components or data, to check their behavior;
 - **Legal and regulatory requirements:** Ensure that tests meet compliance requirements - bias testing, data security and exfiltration tests, tests related to derivative content creation, etc.
- **Governance in the testing process**
 - **Definition of a test plan:** Detailed document describing the scope, objectives, strategies, resources, schedule and acceptance/denial criteria of the test;
 - **Test environments:** Use of separate environments for development, testing and production;
 - **Automation:** Create automated tests wherever possible to ensure consistency and repeatability;
 - **Documentation of results:** Record all test results, including failures and points for improvement, as well as the metrics obtained.
 - **Traceability:** Connecting test results to identified requirements and risks.
 - **Review and approval:** Formal process for reviewing the results and approving the move to the next stage/phase of model development/implementation.

- **Types of tests**

- **FUNCTIONAL TESTS** - These are traditional software development tests, adapted to the AI context. They check that each part of the system behaves as expected.
 - **Unit Testing:** Validating individual components of the AI pipeline. It prevents simple errors from propagating to later stages of the pipeline.
 - **Integration Testing:** Evaluating the interaction between two or more system components.
 - **Validation / Acceptance Test:** Checking that the system as a whole meets the defined requirements and purposes. In this case, it seeks to assess whether the system's functional requirements (what it should do) and business requirements (e.g. reducing defaults) are being met. The most common method is comparison with benchmarks and performance expectations vs. previous results.
- **PERFORMANCE TESTS** - Evaluate the effectiveness of the model under different conditions. They are especially critical in real-time applications.
 - **Latency Test:** Response time for a single prediction. For example, an AI system for medical diagnosis, in order not to affect the queue, must generate a response in less than 1 second.
 - **Throughput test:** Number of predictions processed per second/minute by the model. It is particularly important in high-volume environments, such as financial systems or media analysis.
 - **Scalability testing:** Evaluating whether the system behaves with an increase in the volume of data or users. This can be done with simulations of increasing load - *stress testing*.
 - **Resource Usage Test:** Evaluate the monitoring of the model's CPU, memory, network and energy consumption. Applicable to embedded models, mobile devices and *edge computing*.
- **ROBUSTNESS TESTS** - Tests that assess the model's **resilience** in the face of non-ideal or unexpected situations.

- **Test with noisy data:** Check that the model maintains its performance with imperfect data - for example, typos, visual noise, etc.
 - **Testing with Outliers and Extreme Values:** Check how the model reacts to rare, extreme or anomalous inputs. Depending on the inclusion of these types of "dirt", the model may simply "break" or generate inaccurate or completely absurd answers.
 - **Data / Concept Drift Test:** Evaluating the system's performance with data that deviates from the distribution already contained in the training dataset. For example, a model trained on data from 2022 could fail in 2025 if social, economic or behavioral patterns change in its application context.
- **SECURITY TESTS - Tests** that assess the vulnerability of models to intentional attacks or malicious exploitation.
- **Evasion Test:** Small, almost imperceptible disturbances in the input that aim to cause catastrophic errors in the model.
 - **Data poisoning test:** Technique of inserting malicious data into the training set to influence the model's behavior.
 - **Model Stealing:** Evaluating whether an attacker can replicate a proprietary model by repeatedly querying it and observing its responses, behavior and pattern. Related to the risk of trade secrets and intellectual property - **the developers of DeepSeek have been accused by OpenAI of Model Stealing.**
 - **Testing for Privacy Attacks: Aims** to hinder *membership inference* (inferring whether a specific piece of data is part of the training set) and *attribute inference* (inferring sensitive attributes of the input data). It also seeks to avoid *exfiltration / regurgitation of personal data* from the models, which could lead to privacy violations.
- **FAIRNESS TESTS - Essential** to ensure that the model does not produce discriminatory or unfair decisions, either directly or indirectly.
- **Group metrics:** seeks to assess whether the model is less accurate for the demographic groups contained in the training data sets - for example, is the model less accurate for black women than for white men?

- **Individual metrics:** Seeks to evaluate a person or a small group of people individually against each other, to assess whether they receive fair treatment. Ensures fairness beyond statistical averages.
- **INTERPRETABILITY / EXPLANABILITY TESTS** - Fundamental for models used in critical contexts, such as health, finance/credit, employability and access to public/essential services.
 - **Qualitative assessment:** Human review of the decisions generated. Check that the explanations are understandable, reliable and minimally transparent to human *stakeholders* (e.g. clients, regulators and auditors), according to each expertise and objective.
 - **Quantitative evaluation:** Evaluate metrics related to:
 - **Fidelity:** Do the explanations really reflect the model's behavior?
 - **Consistency:** Are the explanations similar for similar inputs?
 - **Stability:** Small changes in the input should not radically alter the explanation of the results obtained.
- **Ensemble methods**

Ensemble is the practice of combining two or more machine learning models **to produce a single prediction or decision**. Instead of betting on just one "champion model", ensemble treats **each model (learner) as a "voter" or "expert" and uses combination rules (vote, average, meta-model etc) to arrive at the final output**.

- **Learner** - Any machine learning algorithm trained on a set of data - for example, a decision tree, a logistic regression or a small neural network. Within **ensembles**, we have **several learners**.
- **Voter** - A learner whose output goes into a voting scheme - majority, average and weighting. Focuses on the mechanics of the decision - each person casts a vote.
- **Expert** - A learner trained to be very good at a sub-task or sub-region of the data space.

In other words, by evaluating these "terms", we can see that all the "voters" and "experts" are learners.

What are the *ensemble* methods?

- **Bagging (*Bootstrap Aggregating*):** several models are trained on slightly different samples from the same dataset and then the guesses are "added together". It serves to increase robustness against fluctuations or noise in the *dataset*.

- **PRACTICAL EXAMPLE:** Random Forest as Ensemble - several *decision trees* are created using different samples from the data set, with *bootstrap* replacement. Each tree votes on its "answer" and the final model chooses either the majority of votes (classification) or the average of the outputs (regression).

The aim here is to reduce OVERFITTING and increase the accuracy and robustness of the model.

- **Boosting:** Each new model receives special focus on the mistakes that the previous one made, like a chain review. It is relevant in situations where a single model has become "myopic" and needs to reduce bias rates, and can improve coverage in minority groups.
- **Stacking:** Heterogeneous set (neural networks, decision trees, regression, etc.) whose outputs become "inputs" for a final model that decides when to trust each one. In other words, models create outputs that serve as "food" for the final model, which decides which set to "feed".
- **Blending:** a more agile form of *Stacking* - a combination of models that uses a small reserved piece of data to train the model that combines everything.

The benefits of using *ensembles* are:

- **Increasing accuracy without overfitting:** By combining modes that see different variations of the data set (*bagging*) or that correct each other's errors (*boosting*), the system converges to a lower overall error.
- **Reduce variance (robustness):** Several independent models neutralize occasional fluctuations in production data, avoiding erratic *edge-case* decisions.
- **Mitigate systematic bias:** *Boosting* forces each learner to focus on poorly rated examples – can potentially improves sensitivity in minority groups, heterogeneous ensembles capture different perspectives.
- **Provide uncertainty estimates:** The dispersion between predictions in the set serves as a proxy for confidence; useful for *reject option* or human review.

- **Resilience to individual failures:** If a model shows *drift* or error, the ensemble's majority vote limits the impact.
- **Regulatory flexibility:** With *stacking/blending* it is possible to add or remove models to suit new requirements (e.g. explainability, latency) without recreating the entire pipeline.

8. Strategic data set division management

An inadequate division of data can lead to optimistic (or pessimistic) estimates of actual performance and promote the occurrence of *overfitting*. Therefore, the way in which the available data is divided up for different purposes during the life cycle is a relevant methodological decision.

- **Key data set**

- **Training Set:** The bulk of the data (usually 60-80%), used to teach the model to learn the patterns and relationships present in the data. The model adjusts its internal parameters (weights, biases) based on this data;
- **Validation Set:** A smaller portion of the data (usually 10-20%), used during the training process to:
 - Adjust model hyperparameters (EX: learning rate, number of layers in neural network) that are not learned directly from the training data;
 - Monitor the model's performance on data not seen during training and detect *overfitting* (when performance on the training set continues to improve, but performance on the validation set begins to deteriorate);
 - Decide when to stop training (*early stopping*);
- **Test Set:** A final portion of the data (usually 10-20%), kept completely separate throughout the training and hyperparameter tuning process. It is used once, at the end, to provide an unbiased estimate of the model's performance on completely new and unseen data. The result on the test set represents the best estimate of how the model will behave in the real world when it is implemented.

- **Strategies for dividing data sets**

- **Simple Random Split:** The data is shuffled and randomly split into the desired proportions. Suitable for large data sets where the distribution is relatively uniform.
 - **Stratified Split:** Ensures that the proportion of different important classes or groups (e.g. product categories, demographic groups) is the same in all three sets (training, validation and testing). Essential for classification problems with unbalanced classes or when fairness between groups is a concern.
 - **Time-based split:** Used when the data has an important chronological order (e.g. time series, log data). Training is done with older data, validation with "intermediate" time data and testing with the most recent data. This realistically simulates how the model will be used to predict the future.
 - **Geographical or entity division:** Used when there are special dependencies or groupings in the data (e.g. separating data by store, by customer, by region) to prevent information about the same entity from leaking between sets.
- **Governance considerations:**
 - **Separate data sets by purpose and define restricted use:** Ensure that no information from the validation or test set is used, directly or indirectly, during model training. This includes not adjusting hyperparameters based on the test set.
 - **Representativeness:** Ensure that all data sets are representative of the target population, especially the test and training set.
 - **Statistical independence:** The sets must be statistically independent for the evaluation to be effective and valid.
 - **Documentation:** Clearly record the division strategy used, the proportions and the reasons for the choice.
 - **Reproducibility:** Use fixed *Random seeds* during splitting to ensure that the same sets can be recreated if necessary in the future.

A practical example would be the creation of an AI system for classifying images of cats and dogs. Using the 70/15/15 stratified split ensures that the proportion of dog and cat images is the same in the training, validation and test sets, preventing the model from being unfairly evaluated if, by chance, the test set contains mainly images of one type.

9. Using synthetic data - uses, benefits and considerations

In many AI development scenarios, obtaining real, sufficient and high-quality data can be a significant challenge. Data can be scarce, expensive, difficult to collect, or contain sensitive information that restricts its use. In such cases, synthetic data - *data artificially generated by algorithms or models to mimic the statistical properties of real data* - emerges as a valuable approach, although its use requires careful governance considerations.

The aim, when creating synthetic data, is for them to share statistical characteristics (distributions and correlations) with the real data they are intended to simulate.

Synthetic data can be generated in different ways, whether (i) based on statistical rules and models, from known distributions, business rules; (ii) based on machine learning models, which can use generative adversarial networks, variational *autoencoders*; (iii) *flow-based* models, through diffusion models and (iv) simulation/agent-based, with the creation of virtual environments or agendas that integrate for data generation.

- **Use cases and benefits**

- **Privacy:** Generating synthetic data that preserves statistical patterns but does not contain actual individual information, allowing data to be shared or used for training without exposing personal data and extinguishing applicable risks of personal data leakage and regurgitation of personal data that would be contained in the training *dataset*.
- **Data Augmentation:** Complement small or unbalanced real data sets, especially for minority classes or rare scenarios.
- **Test and development:** Create realistic data sets to test systems or train models without relying on production data.
- **Scenario simulation:** Generate data for scenarios that are difficult or dangerous to replicate in the real world (e.g. car accidents to train autonomous vehicles);
- **Data sharing:** Facilitate collaboration and research by allowing the sharing of synthetic data when real data cannot be shared.

- **Synthetic data quality analysis**

- **Statistical fidelity:** How well does the synthetic data replicate the statistical properties (marginal distributions, correlations, etc.) of the synthetic data replicate the statistical properties (marginal distributions, correlations, etc.) of the real data? Statistical metrics and comparative visualizations can be used.

- **Utility:** How well does a model trained only with synthetic data (or a combination) compare to a model trained only with real data for the target task? Evaluating the model's performance is the ultimate test of utility.
- **Privacy:** How well does synthetic data protect the privacy of real records? Tests such as *membership inference attacks* can assess the residual risk.
- **Governance and risk considerations**
 - **Biases:** Synthetic data can inherit or even amplify biases present in real data. *Fairness* assessment is also necessary;
 - **Fidelity vs. Privacy:** There is an inherent trade-off: more faithful synthetic data can retain more information about the real data, consequently increasing privacy-related risks.
 - **Validation:** It is essential to rigorously validate the quality and usefulness of synthetic data before using it to train critical models.
 - **Transparency and documentation:** Clearly document when and how synthetic data was used, including methods of generation and evaluation.
 - **Limitations:** Synthetic data may not capture complex relationships or rare *outliers* present in real data.

10. Monitoring and risk mitigation during model training and testing

The process of training and testing AI models also requires risk assessments and controls. Various problems can arise, compromising the performance, *fairness* or safety of the resulting model. Effective governance requires the implementation of continuous monitoring mechanisms to detect these risks early and the application of appropriate mitigation techniques.

- **Common risks**
 - **Overfitting:** The model learns the training data excessively well, including noise and sample-specific patterns, but fails to generalize to new and unseen data. In other words, it is so adherent to the training that, when run in practice with new data, it performs poorly on the validation and test set.
 - **Underfitting:** The model is too simple to capture the complexity and underlying patterns of the data. This results in poor performance on both the training and validation/test sets.

- **Algorithmic Bias:** The model produces systematically unfair or inaccurate results for certain subgroups of the population. Biases can arise in the training data, the choice of features, the algorithm or the optimization criteria.
 - **Data Poisoning:** A malicious attack where an adversary injects corrupted or misleading data into the training set in order to manipulate the model's behavior or degrade its performance.
 - **Misuse of datasets for different purposes / Data Leakage:** Inadvertent inclusion of test or future dataset information in the training set, leading to an unrealistically optimistic performance evaluation.
 - **Training instability:** The training process can fail to converge, oscillate or be very sensitive to small changes in the data or hyperparameters.
 - **Concept / Data drift:** Changes in the statistical distribution of the data or in the relationship between features and the target variable over time, causing a model trained on older data to lose performance.
- **Monitoring strategies:**
 - **Learning curves:** Plot the model's performance (e.g. error, accuracy) on the training and validation sets over the training epochs. A large gap between the curves indicates *overfitting* - both curves stagnating at low performance indicate *underfitting*.
 - **Performance metrics:** Monitor relevant metrics (precision, recall, F1, AUC etc) in the validation and test sets.
 - **Fairness metrics:** Calculate and monitor **fairness** metrics (demographic parity, equal opportunities, etc.) between different subgroups during and after training the model;
 - **Monitoring data distribution:** Compare the statistical distribution of the training, validation and test data to detect any *drift*.
 - **Error analysis:** Investigate the errors made by the model to identify patterns or subgroups where it fails the most.
 - **Visualization of activations/weights:** In *Deep Learning*, checking the internal activations and weights of the model can help diagnose problems.

Finally, the monitoring and mitigation process should be documented, with clear *thresholds* for intervention. Decisions on which mitigation techniques to use and their *trade-offs* (e.g. mitigation of bias by slightly reducing overall accuracy) should be recorded.

- **Lines of defense in terms of ownership and risk management**

- **First line of defense:** Business/product/data engineering teams (system owners and maintainers). They have the role of creating, operating and owning risks, as well as implementing day-to-day controls.
- **Second line of defense:** Risk, Compliance, Privacy, AI Governance teams. These are the governance and risk analysis teams. These teams define frameworks and monitor whether the first line of defense follows the policies and recommendations.
- **Third line of defense:** The organization's internal audit team. Provides independent assurance that the first and second lines are functioning properly and fulfilling their duties.
- **Fourth line of defense (optional):** External consultancy / audit or regulatory action through the regulator. It monitors regulatory compliance, assessing compliance with standards and rules.

III- Governance continues post-development

1- Assessing Release Readiness

Before deploying an AI model/system in a production environment, where its decisions can have real consequences, it is imperative to carry out a final, comprehensive assessment of its readiness.

This **gatekeeping** process ensures that the model meets all the established technical, legal, ethical and business requirements, and that the residual risks are acceptable and manageable. Governance at this stage involves defining clear acceptance criteria and a formal review and approval process.

- **Evaluation frameworks**

To ensure a systematic evaluation, organizations can use:

- **Pre-Release Checklists:** Detailed lists covering areas such as performance (final metrics in the test suite), robustness (stress and adversarial test results), fairness (bias audit completed), security (vulnerability tests performed), legal compliance (compliance with specific regulations), documentation (*model card*, complete technical documentation), monitoring plan, *rollback* plan and user/operator training.

- **Risk/readiness scorecards:** Quantitative or qualitative assessment of the status in each critical area, resulting in an overall readiness score or residual risk profile.
- **Stage-Gate Process:** Definition of a formal decision point in the project lifecycle, where a multidisciplinary Governance Committee reviews all the evidence of readiness and decides whether to *go with conditions* or *no-go*.
- **Key acceptance criteria**
 - **Technical performance: does** the model meet the minimum performance thresholds?
 - **Robustness and security:** Has the model shown adequate resilience in robustness and security tests? Have the vulnerabilities identified been corrected or mitigated?
 - **Fairness and Ethics:** Have the risks of bias been adequately mitigated? Does the model operate in accordance with the ethical principles of the organization? Are the results of the *fairness* audit acceptable?
 - **Legal and regulatory compliance:** Have all legal and regulatory requirements applicable to the project been met? Is the necessary documentation complete and up-to-date?
 - **Operationalization:** Is the implementation plan clear? Are the monitoring mechanisms in place? Is the operations team trained? Is there a contingency/rollback plan?

It may therefore be advisable to set up a multidisciplinary committee (including representatives from different areas of the company), which will be responsible for the final decision. All the supporting documentation must be presented to this committee, which will formally decide whether or not to release the model, in a reasoned manner, including any conditions imposed.

2- Creating and using *Model Cards* effectively

Model Cards have emerged as a tool to promote transparency and accountability in the development and deployment of AI models. Inspired by nutrition labels and fact sheets, they **provide a structured and accessible summary of the most important information about a specific model, its data, its performance, its limitations and its ethical considerations**. The creation of a comprehensive Model Card is often associated as a key requirement in the assessment of readiness to release a system for deployment.

The information contained in a Model Card can be used by stakeholders, risk specialists, regulators and end users, allowing them to better understand how the model works and to make informed decisions about its use. The level of detail and language may need to be adapted depending on the target audience.

- **Essential components of a Model Card**

- **Basic information:** model name, version, date, developers/organization responsible, overview of its purpose and intended use cases; information about its license (if applicable).
- **Usage considerations:**
 - **Primary use cases** (what are the specific tasks that the model has been designed and validated to perform?);
 - **Out-of-scope use cases:** For which applications is the model unsuitable or untested?
- **Limitations:** What are the known weaknesses/limitations of the model? Under what conditions can it fail or have its performance degraded?
- **Risks and biases:** What are the potential biases identified in the model's behavior and what are the potential ethical, social and security risks regarding its execution.
- **Data:** Description of the data sources used in training and validation. Processing and pre-processing applied to the data. Demographics and characteristics of the *datasets* (particularly relevant for *fairness* purposes). What are the known limitations of the *datasets* used (e.g. low representativeness of certain groups).
- **Evaluation:** Performance metrics used and the quantitative results of the evaluation on the test set. Disaggregation of performance by relevant subgroups (*fairness* analysis). Results of robustness or security tests.
- **Additional information:** Details on the architecture of the model, the training algorithm, hyperparameters and information on the computing environment.
- **Contact information:** Indication of an e-mail channel to obtain more information (if applicable) and/or report problems.

The process of creating a Model Card must be done in collaboration between the development, product, risk and ethics teams. All statements in the Model Card must be supported by documented evidence (test and evaluation results).

It is created during the development of the model and is updated as new information becomes available. In addition, it is also recommended that it be versioned and updated to reflect new changes in operation, risks and considerations to be made about the model.

As for its availability, it is recommended that it be made accessible to the relevant stakeholders and, in the case of commercial licensing of the model to third parties, that it be sent for the purposes of complying with regulatory obligations and for control/audit purposes by the organization that licensed the system.

What is the difference between a Model Card and a System Card?

Both are documentations that promote transparency and accountability in the development and deployment of AI systems. The main difference between the two lies in the scope of the documentation.

The **Model Card** focuses on the individual model. The **System Card** focuses on the complete AI system.

- **Model Card:** Focuses on a specific machine learning model. It details information such as the model's architecture, training data, performance metrics, limitations, potential biases and intended use. It is the **NUTRITIONAL LABEL** of an AI model.

- **System Card:** This has a broader scope, addressing the AI system as a whole, which can include multiple models, software components, hardware and even policies and processes that govern its use. It describes how the system works at a higher level, its capabilities, possible interactions and the mitigations put in place for potential risks and damage.

3- Conducting pre-deployment pilots

Before a large-scale deployment of an AI system, especially those with significant impact potential or high risks, conducting a controlled pre-deployment pilot is a prudent and highly recommended governance practice.

A pilot aims to test the system in a real or semi-real environment, but on a limited basis, to collect data on its performance, *feedback* from users, identify unforeseen problems, validate assumptions and refine its deployment strategy before full launch.

- **Methodology**

- **Planning:** Clearly define the specific objectives of the pilot and its scope (which functionalities, which users/segments, which environment), the success metrics, its duration and the resources needed to carry it out.

- **Selection:** Carefully choosing the participants (users, customers) or the environment (internal - within the organization or external - for real customers), ensuring that they are representative but manageable.
 - **Controlled implementation:** Closely monitoring the implementation, curating and closely monitoring its performance.
 - **Data collection and feedback:** Systematically collect quantitative data (performance metrics and logs) and qualitative data (interviews, questionnaires, observations), especially from users, during the pilot.
 - **Analyze the data:** Analyze the data and *feedback* collected, in relation to the defined objectives and metrics. Identify successes, failures, lessons learned and areas for improvement.
 - **Go / no-go / adjust decision:** Based on the results, make a decision on whether to proceed with the full deployment, make adjustments to the system or the deployment strategy, or even cancel the system launch if significant problems have been discovered.
- **Types of pilots**
 - **Shadow mode:** AI system operates in parallel with an existing process, generating predictions and recommendations, but without these being used to make actual decisions. Allows AI performance to be compared with the *status quo* without operational risk.
 - **Canary Release:** System released gradually to a small percentage of users, which is gradually increased while performance and stability are monitored. Allows problems with limited impact to be assessed.
 - **A/B Testing:** Two or more groups of users are exposed to different versions (e.g. old system vs. new AI system) and their results are statistically compared to determine which one performed better.
 - **Geographical / departmental pilot:** The system is initially deployed in a specific geographical location or internal department of the organization before being launched.

4- Establishing maintenance and upgrade schedules

The implementation of an AI model is not the end of the line for governance - it is the beginning of its operational life, which requires continuous maintenance and monitoring, as well as updates and often periodic training to ensure that it continues to operate effectively, safely and relevantly.

Governance at this stage involves establishing a clear timetable and processes for this activity, ensuring that they are carried out in a proactive and controlled manner.

- **Why is ongoing maintenance necessary?**

- **Data drift and concept:** The real world changes. The statistical distribution of input data or the relationship between *inputs* and the desired *output* can change over time, degrading the performance of a model trained on older data.
- **New patterns:** New behaviors, threats or patterns can emerge, requiring the model to be updated to recognize them.
- **Feedback and errors:** Continuous monitoring can reveal errors, biases or areas where the model is underperforming and needs to be corrected.
- **Technical improvements:** New algorithms, training techniques or architectures may become available, offering new opportunities to improve the performance and stability of the mode.
- **Changes in requirements:** Business, legal or ethical requirements may change, requiring adjustments to the model or its controls.

- **Components of a maintenance and upgrade plan**

- **Continuous monitoring:** Definition of key metrics (performance, drift, *fairness*, resource usage) to be monitored continuously in the production environment, with thresholds and alerts to indicate the need for intervention.
- **Preventive maintenance:** Regularly scheduled activities to ensure the health of the system (e.g. checking logs, optimizing performance, backups).
- **Periodic Re-Training:** Schedule set to re-train the model with more recent data. The frequency depends on the volatility of the environment and the observed *drift* rate (it can be daily, biannually, monthly, quarterly, etc.).
- **Model / feature updates:** Process for introducing new versions and features to the model.
- **Corrective maintenance:** Procedures for responding quickly to incidents, bugs or problems in the model that are affecting its performance and other key metrics.

- **Maintenance windows:** Planning specific periods to carry out updates that may require the system to be unavailable, minimizing the impact on users.
- **End of life:** Criteria for defining when a model should be decommissioned.
- **Factors influencing frequency and strategy:**
 - **System criticality:** the more critical/higher risk the system, the more intensive the monitoring and updates must be;
 - **Dynamics of the environment:** Rapidly changing environments (e.g. financial markets, social networks) require more frequent updates;
 - **Cost vs. Benefit:** Balance the cost of re-training/updating with the expected benefit in terms of performance and risk mitigation.
 - **Data availability:** The frequency with which relevant new data becomes available.
 - **Team capacity:** Resources available to carry out system maintenance and update activities.
- **Process governance**
 - **Responsibility:** Clearly define who is responsible for monitoring the model, maintaining it and approving updates.
 - **Change Management Process:** Develop a formal process for reviewing, approving and documenting all significant changes to the system.
 - **Regression testing:** Ensuring that updates do not introduce new problems or degrade performance in areas that worked previously.
 - **Communication:** Inform relevant stakeholders of maintenance schedules and significant updates.

5- Carrying out continuous security audits and tests

New vulnerabilities can be discovered after the system has been deployed/released. In addition, threats can evolve and the system's own behavior can change unexpectedly. Therefore, continuous governance requires periodic security audits and tests to verify that the system continues to operate securely, effectively and in compliance with internal policies and requirements.

- **Ongoing audits:** Audits are systematic and independent examinations to assess the adequacy and effectiveness of the controls, policies and procedures related to the AI system.
 - **Internal audit:** Carried out by a specialized team from within the company.
 - **External audit:** Conducted by an independent entity (auditing firm, specialized consultancy).
 - **Regulatory audit:** Required by a specific regulatory body.
 - **Compliance audit:** Focused on adherence to laws, regulations and international standards/good practices (ISO/IEC, IEEE, etc.).
 - **Performance audit:** Evaluates whether the model continues to achieve expected performance metrics;
 - **Fairness Audit:** Periodically evaluates the fairness of the model in relation to different subgroups.

The frequency of the audit can vary according to the risk level of the system, regulatory requirements and the results of previous audits (it can be annual, biannual or more frequent for critical systems).

- **Continuous safety tests:**
 - **PenTesting:** Controlled and authorized attempts to exploit vulnerabilities in the AI system and associated infrastructure by simulating real attacks.
 - **Vulnerability scanning:** Use of automated tools to identify known vulnerabilities in software, libraries or configurations.
 - **Source Code Review:** Periodic analysis of the system's source code to identify possible security flaws.
 - **Adversarial tests:** Continuous evaluation of the model's robustness against evasion, poisoning or other specific attacks.
 - **Threat Modeling:** Periodic reassessment of the system's potential threats considering new attack tactics or changes in the environment.
- **Governance practices**
 - **Planning:** In addition to audits, include in the schedule/plan specific technical tests to be conducted periodically to identify and mitigate security vulnerabilities.

- **Independence:** Ensuring the independence of auditors (internal or external).
- **Documentation:** Record the plans, scopes, methodologies, results and recommendations of all audits and tests.
- **Remediation:** Establish a clear process to track and guarantee the implementation of recommended corrective actions.
- **Reporting:** Communicate the results of audits and tests to stakeholders relevant to leadership.

6- Red teaming and threat modeling

In addition to security audits and tests, two more proactive techniques focused on simulating real adversaries are: Red Teaming and Threat Modeling. They help identify vulnerabilities and risks that may not be apparent through more conventional approaches.

- **Red Teaming for AI:** Inspired by cybersecurity exercises, Red Teaming for AI involves a dedicated team simulating attacks by adversaries with the aim of finding weaknesses in the AI system, the data or the associated processes.

The aim is not only to identify vulnerabilities, but also to assess the effectiveness of defenses, detection and response processes and the overall resilience of the system. The Red Team process can be carried out in the following steps:

- **Planning:** Define the objectives (e.g. trick the model, extract data, cause denial of service), the scope (which components of the system), the rules of engagement (what is allowed to be done) and the composition of the team (expertise in AI, security, ethics, privacy, etc.).
- **Reconnaissance:** Collect information about the target system (architecture, data, APIs, public documentation).
- **Attack development:** Create strategies and tools to exploit potential AI system-specific vulnerabilities (evasion, poisoning, extraction, etc.) or traditional software vulnerabilities.
- **Execution:** Carry out attacks in a controlled manner.
- **Analysis and reporting:** Document successful attacks, exploited vulnerabilities, the effectiveness of defenses and provide detailed recommendations for mitigation.

As such, Red Teams need to have knowledge of specific Machine Learning attack vectors, such as:

- (I) **Evasion attacks:** Creating subtly modified inputs (images and text) that are classified as incorrect by the model;
- (II) **Poisoning Attacks:** Inserting malicious data into the training set (during initial training or through feedback loops) to compromise the model;
- (III) **Model Extraction Attacks:** Inferring the architecture or parameters of the model through repeated queries.
- (IV) **Privacy Inference Attacks:** Attempting to extract sensitive information about the training data from the model by trying to obtain personal databases.
- (V) **Prompt Injection Attacks:** Manipulate prompts to make language models generate unwanted content or reveal confidential information.

➤ **Threat Modeling:** A structured process for identifying, analyzing and prioritizing identified threats. It helps you think like an attacker and anticipate specific risks. The steps in the process are:

- **Asset identification:** Determine which system components are critical and should be protected, such as AI models, training data and user interfaces;
- **Creating architecture diagrams:** Map out the structure of the system, including data flows, entry points and interactions between components;
- **Identifying threats:** Use frameworks such as STRIDE (Spoofing, Tampering, Repudiation, Information Disclosure, Denial of Service, Elevation of Privilege) to categorize possible threats.
- **Risk assessment:** Analyze the probability and impact of each identified threat.
- **Implementing countermeasures:** Developing and applying solutions to mitigate or eliminate threats such as data validation, robust authentication and continuous monitoring.

Red Team vs. Threat Modeling - What are the differences?

The difference lies in when they are applied and the purpose of each approach.

Threat Modeling: This is a proactive approach that aims to identify and mitigate possible threats during the early stages of system development. It involves analyzing the architecture, data flows and possible attack vectors to anticipate vulnerabilities before the system goes into final operation. It involves theoretical analysis and modeling.

Red Teaming: This is a practical, offensive approach that simulates real attacks to test the effectiveness of an organization's defences. In other words, it is practical and reactive, focusing on identifying flaws that can be exploited by real attackers in already operational environments, applying offensive tests.

7- Managing incidents and operational risks

Despite all prevention and testing efforts, incidents and failures can occur when the AI system is operating in the real world. These can be technical failures, performance degradation, unexpected or biased behavior, security or privacy breaches and unintended consequences.

A robust incident and risk management plan is therefore an essential component of ongoing governance, enabling the organization to respond quickly, effectively and in a coordinated manner to minimize damage, restore operations and learn from experience.

➤ **Incident management framework (e.g. NIST SP 800-61):** A structured incident management process involves:

- **Preparation:** The proactive phase. It includes the development of incident response policies and procedures, the formation and training of incident response teams (CSIRTs), the implementation of monitoring and alerting tools and the carrying out of simulated exercises.
- **Detection and analysis:** Identify that an incident has occurred (through monitoring, user reports and system alerts), validate the occurrence, assess the scope and initial impact, prioritize the response.
- **Containment:** Taking immediate action to limit the spread of the damage and preserve evidence. For AI, this may involve temporarily disabling the system (*Kill Switch*), reverting the system to a previous version, isolating affected components or implementing manual controls.
- **Eradication:** Identifying and eliminating the root cause of the incident. This may involve bug fixing, security re-training.

- **Recovery:** Restore affected systems to normal operation safely, validate that the root cause has been eliminated and monitor closely to ensure that the incident does not reoccur.
- **Lessons learned:** Carry out a detailed analysis of the incident, documenting what happened, how the response was managed, what the root causes were and what improvements can be made to controls, processes or training to prevent future cases.

In addition, like other aspects of AI governance, documentation is also a relevant action for incident management, response and containment, and there should be a detailed record of all incidents that have occurred, including *timestamps*, description, impact, actions taken, resolution and lessons learned.

No less important, the organization must define the protocols for communicating this incident both to the responsible and applicable authorities/agencies, but also to the impacted owners/users (if applicable), in addition to internal communication to the teams. This can include senior management, Legal, Compliance, Public Relations and Communications, among other teams.

8- Implementing Effective Operational Controls

To ensure that an AI system operates consistently, safely and in line with the policies and objectives defined on a day-to-day basis, it is necessary to implement a set of operational controls. These are systematic mechanisms and procedures focused on the ongoing management and supervision of the system in production. Governance must ensure that these controls are adequate, implemented correctly and monitored for their effectiveness.

➤ Types of controls

- **Preventive:** Designed to prevent errors, fraud or unwanted events from happening (e.g. input validation, access control, segregation of duties, training);
- **Detective:** Designed to identify errors or faults after they have occurred (e.g. log monitoring, audits, performance alerts);
- **Corrective:** Designed to remedy problems and incidents that have already been detected (e.g. backup and recovery procedures, incident response plans, error correction);
- **Directive:** Designed to guide behavior in the desired direction (e.g. policies, procedures, guidelines and training).

➤ Governance in the implementation and maintenance of controls

- **Planning:** Ensuring that controls are planned/designed in such a way as to be effective in mitigating the risks identified and appropriate to the operational context.
- **Implementation and testing:** Checking that the controls have been implemented correctly and testing their effectiveness periodically.
- **Monitoring:** Continuously evaluating their effectiveness and whether they are working as expected;
- **Continuous improvement:** Adjusting and improving controls based on the results of monitoring, audits and incident analysis.

9- Ensuring public disclosure and technical documentation

Transparency is a fundamental principle of responsible AI governance. Organizations that develop and deploy AI systems have a responsibility to be transparent about their capabilities, limitations and potential impacts. This involves both the public disclosure of relevant information and the maintenance of detailed technical documentation, meeting the expectations not only of users and different audiences, but also of regulators.

➤ Objectives of dissemination and documentation:

- **Building trust:** Transparency helps to build greater trust with users, customers, regulators and the general public;
- **Enable technical analysis:** Provides information for external experts, academics and civil society to evaluate and audit systems;
- **User understanding:** Helps users understand how the system works, how to interact with it and what its limitations and associated risks are;
- **Comply with regulatory requirements:** Some standards (such as the EU AI ACT) require specific levels of transparency and documentation, especially for high-risk systems and when the system interacts directly with individuals.
- **Facilitating accountability:** Detailed documentation is essential for tracking decisions and assigning responsibilities.
- **Support maintenance and future development:** Technical documentation is crucial for the teams that maintain and update the system.

Furthermore, the language should be adapted to each audience. For example, for the general public and end users, considering the complexity of AI systems and the general population's low knowledge of technical aspects and terms, it is recommended to communicate using clear, concise and non-technical language, focusing on the purpose of the system, how it can affect the user and what data is used, as well as the expected benefits and risks. Openness to feedback is also recommended. For regulators and auditors, on the other hand, technical details about the system's operation and performance, impact assessment results and evidence of compliance are expected. Finally, for the technical community, including researchers and academics, organizations can focus on disseminating *white papers*, API documentation, source code (if open-source), detailed model cards and *benchmarks*.

As for the components of this technical documentation, after the system has been implemented, the ongoing documentation should include:

- (i) **Version history:** Record of all versions of the model deployed, with dates and description of changes;
- (ii) **Monitoring results:** Periodic reports on performance, *drift*, *fairness* etc.
- (iii) **Operation logs:** Detailed record of system activity;
- (iv) **Maintenance log:** Documentation of all maintenance activities, enhancements, re-training and system updates.
- (v) **Audit and test reports:** Results of internal and external audits, as well as periodic security tests.
- (vi) **Recording incidents:** Documentation of incidents that have occurred, their causes, resolutions and the consequences (actions taken and lessons learned).
- (vii) **Operational documentation (*Rubbooks*):** Guides for operating, monitoring and troubleshooting the system.

Organizations must also take into account the **balance of transparency, assessing the expected balance between full transparency and possible impacts**. For example, by adopting total transparency in documentation about the AI system, its infrastructure, operation and parameters, organizations can expose Intellectual Property (trade secrets, proprietary algorithms) or create attack vectors (reveal details and flaws to be exploited) by malicious third parties/attackers. Governance must define the appropriate level of transparency for each type of information and audience, justifying its decisions not to disclose certain details on the basis of legitimate risks.

10- Planning system decommissioning and *end-of-life*

Like any technology, AI systems also have a finite life cycle. Eventually, a model can become obsolete or outdated, ineffective or no longer aligned with the business strategy.

Planning the controlled and responsible *decommissioning* and end-of-life of the system is the final stage of AI lifecycle governance.

➤ **Triggers for deactivation**

- **Technical obsolescence:** The underlying model or technology has become outdated, superseded by more effective approaches.
- **Degraded performance:** The model no longer meets the performance requirements due to significant drift or changes in the environment that can no longer be corrected with retraining.
- **Change in business strategy:** The original use case is no longer a priority, or the organization decides to go in a different direction.
- **Maintenance cost:** The cost of maintaining and updating the model becomes prohibitive or no longer justifiable for the value it delivers.
- **Unacceptable risks:** Ethical, legal or security risks are identified and cannot be adequately mitigated.
- **Replacement:** A new system is developed to replace the old one.
- **Regulatory requirements:** New laws or regulations make the existing model no longer compliant, and the organization must choose to decommission it.

➤ **Controlled deactivation process**

- **Planning:** Define a schedule, identify all the components to be decommissioned (models, APIs, data pipeline, infrastructure);
- **Communication:** Notify all affected stakeholders (users, internal teams, partners/suppliers) in advance of the shutdown and the timetable.
- **Transition/Migration:** If a replacement system is being implemented, manage the migration of users and data in a gradual and controlled manner.
- **Archiving:** Securely preserving the model's final code, associated training data, complete lifecycle documentation, relevant logs and historical performance records. This is crucial for future audits, regulatory compliance and organizational learning.

- **Technical decommissioning:** Remove the model from the production environment, deactivate APIs, disconnect the associated infrastructure and revoke access.
- **Validation:** Confirm that all components have been deactivated correctly and that there are no more active dependencies.

DOMAIN IV

Understand how to govern AI deployment and use

I- Understanding the context of the AI use case

1- Understanding the context of the AI use case

The first and perhaps most crucial step in the pre-deployment evaluation of an AI system is a thorough understanding of the specific context of the *use case*.

➤ Business objectives:

It must be assessed what specific business problem the AI wants to solve or what opportunity it seeks to exploit.

The implementation of the system must be directly linked to measurable strategic goals.

EXAMPLE: A bank that implements an AI model for fraud detection aims to reduce financial losses and increase the security of transactions.

Clear objectives are essential for defining success metrics and *justifying* the *business case*.

➤ Performance requirements

Evaluate which metrics - accuracy, latency, scalability, reliability - define success for the model.

A medical diagnostic system, for example, requires a very high level of precision, while a *chatbot* for customer service can understand that the requirements for speed of response are more critical.

Defining these requirements is vital for model selection, training and monitoring.

EXAMPLE: An algorithmic trading system in the financial market needs extremely low latency, while a demand forecasting system can tolerate longer processing times.

When we see a drop in the model's performance, we say that we are dealing with a **MODEL DECAY** - a gradual performance drop, typically driven by model - or data-drift, overfitting, or underfitting.

➤ Categories of low-performance model types

- **High Variance:** The model over-learns the details of the training set, even the noises and exceptions. In other words, it does very well on training, but very badly on new data. The common cause is **OVERFITTING - the model is memorizing rather than actually learning general patterns**.
- **High Bias:** The model doesn't learn well enough - it makes very simple assumptions. In other words, it gets a lot wrong in training and with new

data. The common cause is that the model is either too simple or the training data set is poor - **UNDERFITTING**.

- **Data Drift:** When new data changes over time - it is no longer similar to the data used for training and, statistically, there have been changes in users' consumption/use patterns. Even a good model starts to make mistakes because the context / reality has changed.
 - **Example:** A bank implemented a trained credit model with a database in 2017 and has been using it ever since. Over the years, the bank has noticed a sharp decline in the model's performance as it has seen people's consumption patterns and defaults increase year on year.

➤ **Data availability and quality**

The fuel for AI is data. It is essential to assess whether the data needed to train and operate the model is available, is of sufficient quality (*complete, accurate, relevant, consistent*) and whether the organization has the rights to use it for the intended purpose (considering concerns related to intellectual property and privacy).

The lack of suitable data or the presence of *bias* in existing data can severely compromise the performance and fairness of the model.

Data governance is crucial here, including the assessment of **lineage and provenance**.

➤ **Workforce Readiness**

Should the organization and its employees be prepared to adopt and interact with the new AI system?

This question involves assessing the need for training, redefining roles and processes, and managing the organization's cultural change. Resilience in adoption or a lack of understanding of how to use the AI system correctly can undermine the expected benefits.

2- **Understanding the differences in AI model types**

Choosing the type of AI model is a technical decision with significant governance implications. Different types of models have different types of characteristics, capabilities, risks and governance requirements.

➤ **Classic vs. Generative:**

Classical (or discriminative) models are trained for specific tasks such as *classification* (e.g. identifying spam emails) or *regression* (e.g. predicting real estate prices). They learn to map inputs to predefined outputs.

Generative models (*GenAI*), such as LLMs (Large Language Models) or *Diffusion* models, are trained on vast data sets to create new and original content (text, images, code, sounds, etc.) that resembles the training data. The governance of generative models presents unique challenges, such as the risk of *hallucinations*, the management of intellectual property of the generated content and the potential use for disinformation.

➤ **Closed source (proprietary) vs. Open Source:**

Proprietary Models are developed and controlled by a specific entity, with generally confidential source code and architecture.

Open-source models have their source code publicly accessible, allowing inspection, modification and distribution.

The choice has an impact on transparency, costs, flexibility, information security and third-party risk management.

For example, open source models often improve transparency and allow on-prem deployment, though they require stronger internal security oversight. However, their use requires greater internal expertise for implementation and security, as well as specific licenses which, in certain situations, need to be evaluated.

➤ **Small models vs. large models**

The size of the model, often measured by the number of parameters, impacts the computational resources required for training and inference, the associated costs and, potentially, the complexity and generalization capacity.

Large models, such as many LLMs, require robust infrastructure (often in the cloud) and can be difficult to interpret (*Explainability*).

Smaller models can be more efficient for deployment on devices with limited resources - ***Edge Computing***.

➤ **Language vs. Multimodal**

Language models focus on the processing and generation of text (***Natural Language Processing - NLP***).

Multimodal models can process and combine text, images, audio, or video - **for example, ChatGPT can evaluate an image and describe it in text.**

Multimodality expands the possible applications, but also increases the complexity of data management and risk assessment, as different types of biases and failures can arise for each of these modalities.

3- Understand the differences in AI deployment options

How an AI model is technically deployed and made available for use is also a critical decision with governance implications.

➤ **Cloud vs. On-Premise vs. Edge:**

- **Cloud:** Deploying the model on a cloud provider's infrastructure (e.g. Azure, AWS) offers scalability, flexibility and access to specialized tools. However, it raises concerns about data security, sovereignty and operating costs (OPEX), as well as always generating a dependency on *cloud* providers for operation. Furthermore, for real-time applications, this can also be a concern.
- **Local (On-Premise):** Hosting the model on your own infrastructure offers greater control over data and security, potentially lower latency, but requires a significant initial investment (CAPEX) in hardware, as well as maintenance expertise.
- **Edge:** Running the model directly on local devices (e.g. machines/computers/notebooks, sensors, smartphones) minimizes latency, improves privacy (data doesn't need to leave the device) and allows the model to operate *offline*. However, its application is limited by the device's computing capacity, which makes updates and centralized monitoring difficult.

The choice depends on the use case - **for example, an AI system for analyzing medical images can be deployed *on-premise* for privacy, control and latency reasons, while a recommendation system on an e-commerce website usually resides in the cloud for scalability.**

➤ **Commercial off-the-shelf (COTS) vs. Custom-built**

Although it is unlikely that an "as is" AI system can be applied to certain use cases without enhancement / adaptation, in simpler or "general" use cases, **Commercial off-the-self (COTS)** systems prove to be of great value to organizations.

In general, these are **ready-to-use** systems, **made by a company and sold as a product. A kind of off-the-shelf software.**

FOR EXAMPLE: A company buys a ready-made, general-purpose anti-fraud AI system that it can implement directly into its operations.

Custom-built systems, on the other hand, are those that are tailor-made to the specific needs of the company. In this case, the organization needs to further refine / adapt the system to its use case and purpose, also taking into account the peculiarities of its operations, infrastructure and capacity. They are generally more expensive and time-consuming to implement, considering customization.

➤ **Use "as is" vs. Fine Tuning vs. RAG vs. Other Techniques**

Rarely is a pre-trained model (especially in foundational models) used directly without some adaptation. As such, using it "As Is" (i.e. the way it was developed by the Provider) is less common for complex applications, as it may not perfectly suit the specific context and proposed use case.

Other actions to improve the capability of the use case model:

- **Fine Tuning:** Re-training the pre-trained model with a smaller, domain-specific data set to specialize it for a specific task. It makes it possible to adapt powerful models (such as LLMs) to tasks specific to the organization - **FOR EXAMPLE:** A chatbot that understands the company's internal terminology.
- **RAG - Retrieval-Augmented Generation:** A technique, especially for LLMs, that combines the generative capabilities of the model with the retrieval of information from an external, up-to-date knowledge base. Before generating a response, the system searches for relevant information in specific documents or databases and provides it to the LLM as context, improving factual accuracy and reducing the risk of hallucinations - **FOR EXAMPLE:** an AI assistant for technical support who consults up-to-date product manuals before responding.

In addition to RAG and Fine Tuning, there are other techniques/forms for adapting or integrating models, such as *Prompt Engineering*, Ensembles - discussed earlier, which consists of putting several models together - etc.

➤ **Decision-making techniques / forecasting possibilities**

- **Greedy Algorithm:** This is a "greedy algorithm", which makes the best possible decision at each local step, hoping that this will lead to the best global solution. In other words, it chooses the path that seems best at the time, without thinking about a future global evaluation.
- **Dynamic Programming:** This is a technique that solves problems by breaking them down into "sub-problems", saving the results to avoid recomputation. After breaking down the problems, it assembles the best complete solution based on these pieces - it usually takes longer to find the best overall answer.
- **Backpropagation:** A process used to train neural networks, in which the output errors are taken into account by the system, which tells the "neurons" where they went wrong in order to **adjust their weights**. It's like saying something like "Hey, that was bad, try adjusting it here to make it better" until the model learns.
- **Monte Carlo Simulation:** A statistical technique that uses random repetitions to simulate possible outcomes. Generally used in reinforcement learning. It's a way of testing various possibilities using random draws or trials - it serves to predict what might happen in uncertain situations. **For example:** Doctor Strange, in the movie Avengers: Endgame, predicted

millions of possibilities and only found one that worked - something similar to what we're trying to evaluate with Monte Carlo Simulation.



II- Carry out key activities to evaluate the AI model - *Due Diligence*

1- Carry out or review an impact assessment on the selected AI model

One of the main actions in AI *Due Diligence*/Assessment is to carry out or review a comprehensive impact assessment.

The aim is to **proactively identify, analyze and document the potential risks and benefits associated with using the AI model in the specific context defined in the use case**. This assessment is not a one-off event, but part of an iterative risk management process.

➤ Types of impact assessment

- **Algorithmic Impact Assessment (AIA):** Focused specifically on the unique risks exacerbated by AI, such as algorithmic biases (*Bias*), lack of explainability (*Explainability*), security risks (which can cause *Adversarial Attacks*) and wider social impacts.

- **Data Protection Impact Assessment (DPIA):** Required by privacy regulations (such as the GDPR and LGPD), it aims to assess personal data processing activities that pose a high risk to the rights and freedoms of data subjects. AI often triggers the need for a DPIA considering the intensive use of data, potentially impactful automated decisions, processing of sensitive personal data, *profiling* and monitoring, as well as chances related to bias/discrimination - potentiated with the processing of personal data.
- **Ethical Impact Assessment (EIA):** Analyzes the ethical implications of the AI system, aligning it with organizational values and universal ethical principles such as *Justice, Autonomy and Non-Maleficence*. It considers the impact on different stakeholders and social groups.
- **Human Rights, Democracy and Rule of Law Assurance (HUDERAF):** is a framework proposed by the Council of Europe to ensure that AI systems comply with human rights, democratic principles and the rule of law. HUDERAF combines human rights due diligence processes with technical governance. Broadly **speaking, it suggests that organizations conduct AI impact assessments focused on human rights, democracy and the rule of law - AI HUMAN RIGHTS IMPACT ASSESSMENT.**

In general, a robust impact assessment should, at the very least:

- (i) **Describe the system and the context:** Detail how the model works, the data used, the purpose (use case) and the deployment environment;
- (ii) **Identify risks and benefits:** Map potential risks (legal, ethical, operational, reputational, security and social) and expected benefits for the organization, individuals and society;
- (iii) **Assess necessity and proportionality:** Justify why the use of the AI system is necessary and whether the risks are proportional to the benefits.
- (iv) **Consult relevant stakeholders:** Engage relevant stakeholders (both internal and external) to gather feedback, recommendations and concerns - **Stakeholders concerns.**
- (v) **Identify mitigation measures:** Propose technical and organizational controls to reduce the identified risks to an acceptable and manageable level.
- (vi) **Document:** Record the entire process, findings and decisions for compliance and *accountability* purposes and future reviews.

In addition, it is possible to evaluate the types of *assessments* that can be carried out on AI systems, both by external teams (specialized consultancies) and by the AI Governance team itself:

- **Product Category:** This is the analysis of how the system will be used, who it is made for, what kind of content it generates and whether it respects the rules of use - *such as the AUP (Acceptable Use Policy)*.
 - **Example:** Evaluating whether a chatbot is aimed at customers and employees, whether it can generate images and whether there are filters on the content.
- **Data Usage:** Focuses on what the system does with the data, such as collecting, using, sharing, storing and protecting it.
 - **Example:** Does the system use personal data? Is third-party data shared? What categories of data are used in the system?
- **Technical Specifications:** **Technical** categories of the system, such as architecture, performance, interoperability, APIs, customization capacity, etc.
 - **Example:** What model is being used? Is the proposed infrastructure similar to what the organization usually adopts?
- **Security and Safety:** Evaluates whether the system is secure against attacks, failures, misuse and produces safe and reliable results.
 - **Examples** Are there security controls in place to protect against data leaks? Can it generate offensive or discriminatory content?

2- Identify laws that apply to the AI model

It is important to identify and understand the complex regulatory landscape that applies to the specific AI model and its use case. Legal compliance is a pillar of responsible governance.

Thus, organizations should map not only the specific regulations/standards on Artificial Intelligence (e.g. EU AI ACT and Japan AI Law), but also identify laws related to it:

- **Privacy and data protection** - GDPR, LGPD, PIPEDA, CCPA, etc;
- **Anti-discrimination laws:** related to employment (EEOC - Equal Employment Opportunity Commission / New York Local Law 144 / Illinois Video Interview Act) and credit (ECOA - Credit Opportunity Act);
- **Intellectual Property Laws:** DMCA - Digital Millennium Copyright Act (DMCA); Lanham Act; Defend Trade Secrets Act (DTSA); EU Regulation 2017/1001; Directive 2019/790 and Directive 2001/29 EC - Copyright etc;

- **Product Liability Laws:** Consumer Product Safety Act (CPSA), Colorado SB 24-205 (*Consumer Protections for Artificial Intelligence*), as well as European legal theories - **Strict Liability; Negligence and Breach of Warranty;**
- **Sectoral regulations: Sectors** such as finance and health (HIPPA) have specific regulations;

3- Identify and understand issues unique to a company implementing its own proprietary model

When a company develops a *proprietary software* model itself, it assumes a set of responsibilities that are often greater and pose more governance challenges than when using third-party solutions.

➤ Greater obligations and responsibilities

The organization becomes directly responsible for the entire lifecycle of the model, from its conception and data collection, to training, validation, testing, deployment, monitoring and results. There is no third party to transfer part of the responsibility through a contract.

- **Regulatory compliance:** The organization acts as a PROVIDER / DEVELOPER under legal regimes, such as the EU AI ACT, with more acute obligations.
- **Legal and reputational risk:** Failures, biases or damage caused by the model fall directly on the organization, increasing exposure to litigation and reputational damage.

➤ Specific challenges

- **Transparency vs. Trade Secrets:** Balancing the need for transparency for regulators and end users with the protection of their intellectual property and the trade secrets invested in developing the model - **Transparency vs. Intellectual Property.** Thus, deciding what to document and disclose becomes a complex strategic decision.
- **Training data governance:** The organization must ensure that the data used to train the model has been obtained legally and ethically, especially if it involves external sources (e.g. *webscraping*) or personal user/owner data. Lies about the provenance of this data are a growing risk and several legal debates are underway about this issue.
- **Internal Resource Management:** Requires significant investment in specialized talent (data scientists, ML engineers, ethics and privacy experts), computing infrastructure, *data centers*, robust internal processes for the entire lifecycle;

- **Independent validation and testing:** Although developed in-house, it may be beneficial (or legally required) to have some form of independent validation or audit to ensure the objectivity and robustness/resilience of the model.
- **Ownership of outputs:** Clearly define the ownership of the *outputs* generated by the proprietary model, especially if it is developed for third parties (via API),

III- Governing the deployment and production use of the AI model

1- Apply policies, procedures, best practices and ethical principles in the *deployment of the model*

The transition of the model from the development/test environment to the production environment (*deployment*) must be actively managed, applying the established governance guidelines.

➤ Data governance in production/deployment

Policies related to data governance, assessing issues related to quality, privacy, security and retention, must be strictly applicable to the data that feeds the model in production and the data generated by it.

This includes ensuring that only necessary data is used - *Data Minimization*, that data security is maintained and that privacy requirements are respected in the operating environment.

➤ Deployment risk management

The risks identified in the previous phases must be actively managed during and after implementation - **Continuous Monitoring**. This may involve implementing technical controls (e.g. firewalls, access controls), operational procedures (e.g. manual reviews of certain outputs) and specific monitoring for the risks identified. **The approach should be risk-based** - prioritizing the most latent risks and those that could have the greatest impact on users and society.

➤ User and stakeholder training

End-users and other stakeholders who will interact with the AI system must receive adequate training and guidance on how to use the AI system correctly, its capabilities and limitations, and how to report problems or concerns. This is part of **Workforce Readiness** and aims to prevent misuse and maximize effectiveness.

➤ **Application of ethical principles**

Ethical principles (fairness, transparency, accountability, etc.) should guide deployment and configuration decisions. For example, setting up appropriate levels of human supervision (***Human-in-the-loop / Human Oversight***) in critical processes or ensuring that user interfaces clearly communicate when a decision is/has been assisted/taken by AI.

2- **Establish continuous monitoring and maintenance / re-training schedule.**

AI models are not static - their performance can degrade over time due to changes in the input data (***Data Drift***) or the operating environment (***Model Drift***). Therefore, continuous monitoring of deployed models is essential to detect these problems early.

➤ **Definition of metrics and *baseline***

The organization must establish **clear metrics to evaluate the model's performance in production - EX: accuracy, false positive/negative rate, latency, *fairness* in different subgroups / *fairness***).

In other words, the organization must define a performance *baseline* at the time of implementation to compare with future measurements.

➤ **Performance monitoring**

Implement tools and processes to continuously track the defined metrics. This can involve automated logs, visualization *dashboards* and alerts for significant deviations (***Deviations, Irregular decisions***).

➤ **Drift monitoring**

Specifically monitoring changes in the statistical distributions of input and output data to detect ***Data Drifts*** or ***Model Drifts***. Specialized tools can be used for this.

➤ **Fairness and bias monitoring**

Continuously evaluate whether the model is producing fair results - ***Fairness*** - for different demographic or protected groups, using appropriate fairness metrics.

➤ **Maintenance and re-training schedule**

Based on the monitoring results and the nature of the use case, establish a regular schedule for maintenance - **e.g. software maintenance, security patches, possible retraining of the model**.

Retraining may be necessary to adapt the model to new patterns in the data or to correct performance degradation. The frequency will therefore depend on the volatility of the environment - **e.g. market prediction models may need to be retrained more frequently than image recognition models, which are usually more stable**.

➤ **Versioning**

Keep clear records of the different versions of the model and the data used to train them. This is crucial for traceability and for reverting to previous versions if necessary.

3- Conduct audits, red teaming, threat modeling and periodic security tests

In addition to continuous monitoring, more in-depth periodic evaluations are needed to ensure the robustness, safety, reliability and resilience of the model in the production/deployment phase.

- **AI audits:** systematic and independent reviews (internal or external) to assess the model's compliance with internal policies, regulations, standards/good practices and ethical principles/fairness. Audits can focus on aspects such as data quality, algorithmic fairness, security, transparency and documentation.
- **Red Teaming:** Exercises in which a team (internal or external) simulates **adversarial attacks** to proactively test the security vulnerabilities and robustness of the model. The aim is to identify flaws before malicious actors take advantage of them. This can include **Data Poisoning**, evasion or model extraction attacks.
- **Threat Modeling:** Development of a structured process to identify possible threats to the security of the AI system, vulnerabilities and necessary mitigation controls. Analyzes the system from an attacker's point of view.
- **Security tests:** Technical evaluations focused on the security of the infrastructure that hosts the model, the access APIs and the model itself against cyber attacks.
- **Robustness assessment:** Testing how the model behaves under unexpected or adverse conditions - **e.g. noisy data, inputs outside the expected distribution.**

4- Prevent secondary uses or downstream damage. Define external communications

Governance of use in production also involves anticipating and mitigating unintended consequences and managing effective communication about the system.

- **Prevention of unwanted secondary uses:** The model has been designed for a specific use case. Controls (technical and political) must be implemented to prevent it from being used for unintended or inappropriate purposes, which could generate new risks.
 - **EXAMPLE:** A facial recognition model developed for secure access control should not be used secondarily for employee surveillance without prior evaluation and adequate justification.
- **Downstream damage mitigation:** Consider how the outputs or decisions of the AI model can impact subsequent processes or decisions (*downstream*) and potentially cause damage. **Cascading risks** need to be mapped and mitigated.

- **EXAMPLE:** An error in an inventory forecasting system can lead to inappropriate purchasing decisions by the organization, resulting in financial losses.
- **External communications and transparency:** The organization must define what, how and when to communicate externally about the AI system. This may include:
 - **User notifications:** Inform **users** when they are interacting with an AI system or when a significant (and automated) decision about them has been assisted/taken by AI, as required by law.
 - **Public disclosures:** Publish information about the model (**e.g.** through Model Cards or summaries of technical documentation) to increase transparency and trust, especially in high-risk systems.
 - **Regulatory communication:** Providing required information to regulators (e.g. logs, technical documentation, test results and incident reports).
 - **Crisis communication / Incident Response:** Have plans in place to communicate externally in the event of incidents or serious/relevant model failures. The communication approach will depend on factors related to regulation, sector, type of system and the public involved/impacted.

5- Create controls to deactivate or isolate (*geofencing*) the model if required

There must be mechanisms to quickly intervene in the AI system in production when necessary.

- **Deactivation mechanisms - kill switch:** Implement technical controls that allow the organization to deactivate the AI model quickly in the event of a serious failure, dangerous behavior, security breach or legal/regulatory requirement. This is especially critical for high-risk systems.
- **Location / restriction controls:** Implement the ability to operate the model in certain geographies (e.g. prohibited countries), user populations (e.g. age control) or specific functionalities, if necessary. This may be required to comply with local regulations or to contain a problem identified in a specific subset of users or data.
- **Intervention procedures:** Clearly define the procedures and authorities responsible for activating these deactivation or localization controls - **e.g., notifying Brazil's telecom regulator (ANATEL) if connectivity controls are breached.**