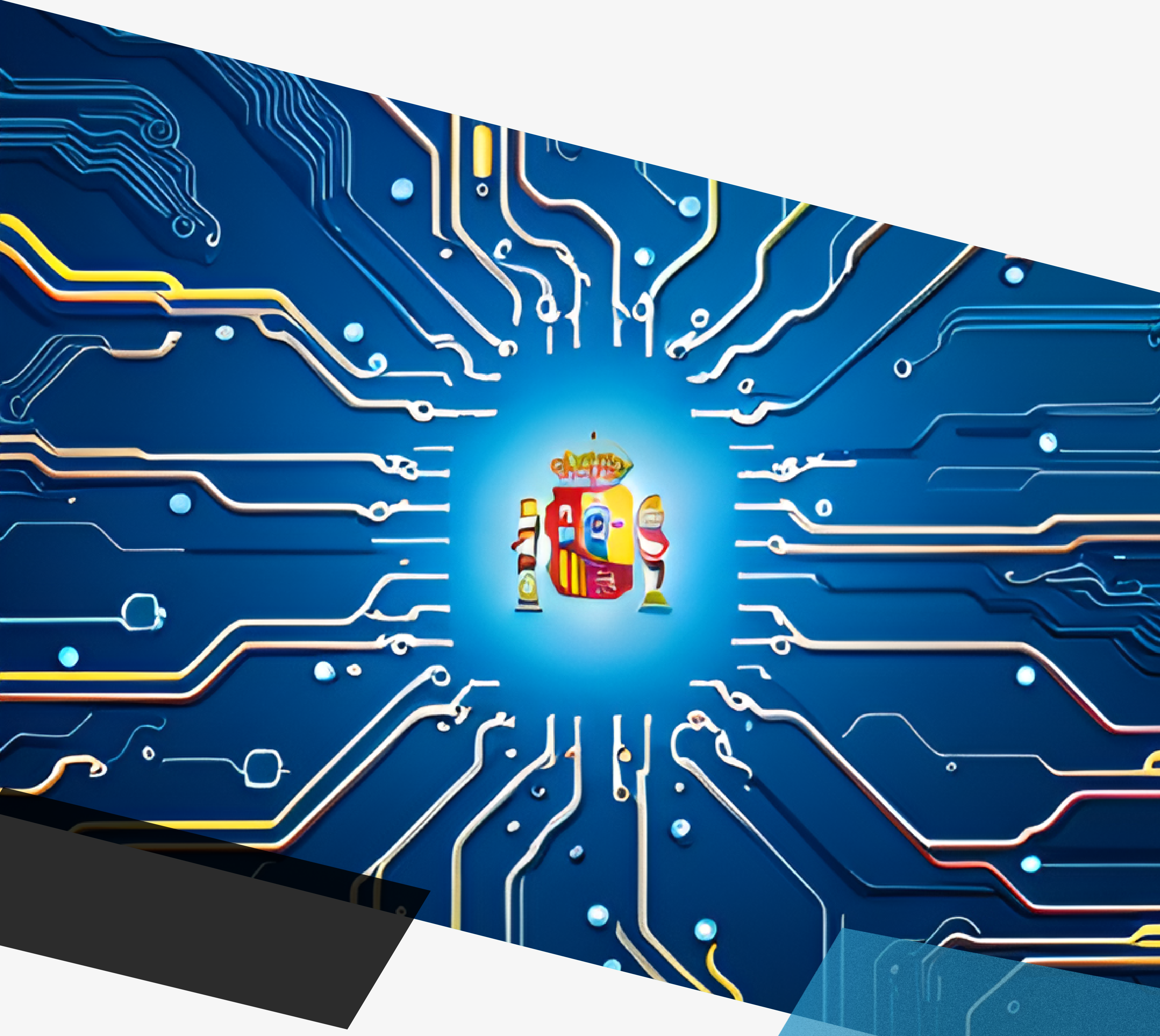


ARTIFICIAL INTELLIGENCE RISK MANAGEMENT IN SPAIN

RIESGOS CATASTRÓFICOS GLOBALES
JUNE 2023

Executive Summary

Artificial Intelligence (AI) is advancing rapidly, bringing significant global risks. To manage these risks, the European Union is preparing a regulatory framework that will be tested for the first time in a sandbox hosted by Spain. In this report, we review the risks that must be considered to effectively govern AI and discuss how the EU AI Act can be implemented effectively.

To facilitate their analysis, we have classified AI-related risks into (i) adversarial risks and (ii) structural risks. The first group includes risks where there is a direct relationship between the harm and its cause. Specifically, we have identified two potential origins: malicious actors with the intention of misusing AI and the AI systems themselves, which can act in unintended ways if not aligned with human objectives. The second group, structural risks, are those caused by the large-scale deployment of AI, focusing on the collateral effects that such technological disruption can have on society.

Regarding adversarial risks, we have focused on three types of threats: (i) cyberattacks and unauthorized access, (ii) development of strategic technologies, and (iii) user manipulation. Cyberattacks and unauthorized access involve using AI to carry out cyber offenses in order to gain resources. The development of strategic technologies refers to the misuse of AI to gain competitive advantages in the military or civilian sphere. User manipulation involves using persuasion techniques or presenting biased or false information through AI.

In terms of structural risks, we focus on five: (i) labor disruption, (ii) economic inequality, (iii) amplification of biases, (iv) epistemic insecurity, and (v) automation of critical decision-making and management processes. Labor disruption entails massive job loss due to automation. Economic inequality suggests that the accumulation of data and computing power might help AI developers concentrate wealth and power. Amplification of biases relates to the biases that algorithms may incorporate and generate in their decision-making. Epistemic insecurity indicates that AI can hinder the distinction between true and false information, affecting a country's socio-political stability and proper functioning. Finally, the automation of critical processes involves handing over command and control of strategic infrastructure to AI, increasing accidental risks.

Once the risks associated with AI are understood, we make nine recommendations to reinforce the implementation of the EU AI Act, especially in view of the development of the sandbox in Spain. The proposals are divided into three categories: measures for the development phase of AI systems, measures for the deployment phase, and conceptual clarifications.

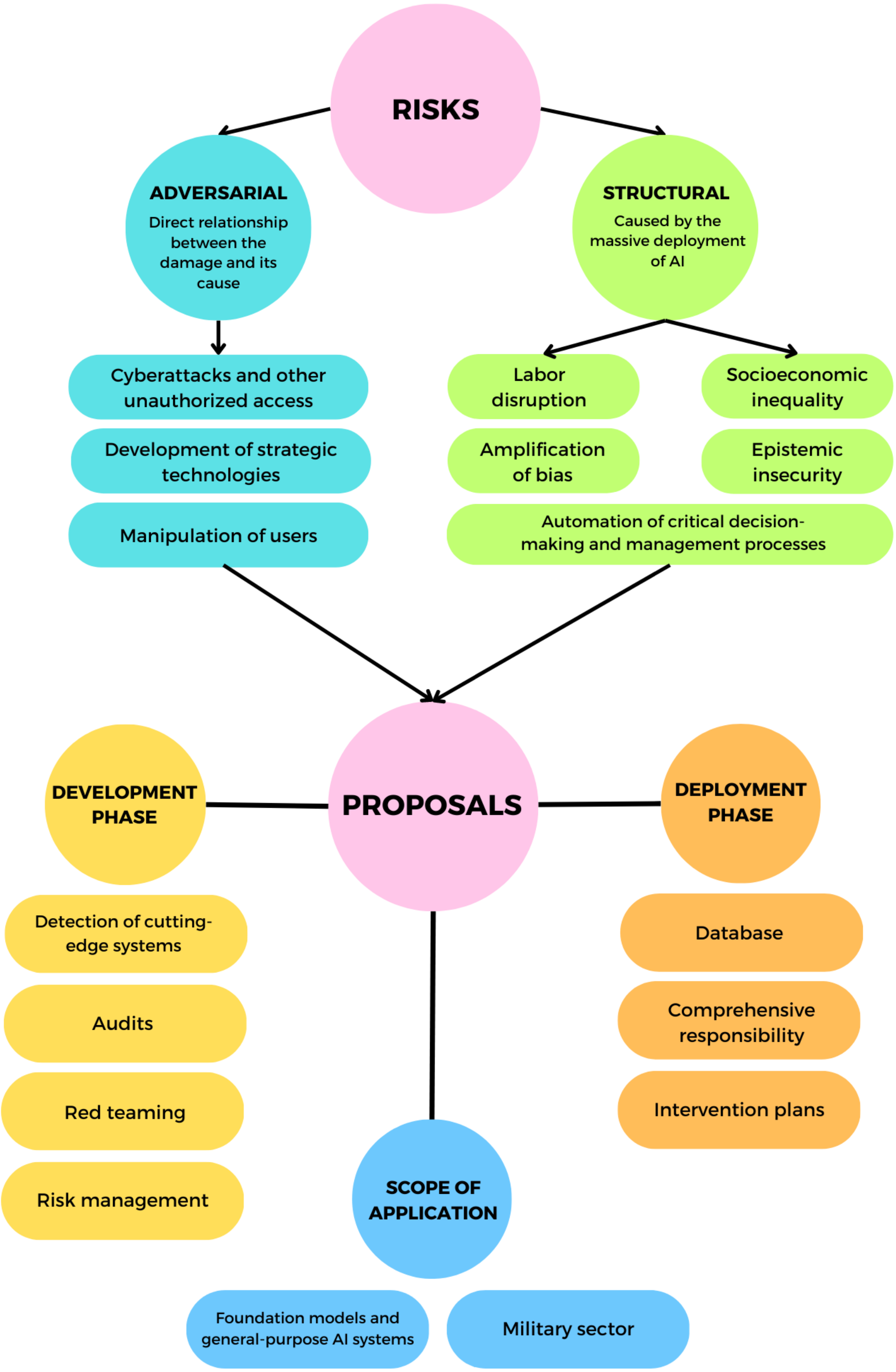
Regarding the measures for the development phase, we prioritize four: (i) detection and governance of cutting-edge systems, taking training compute as an indicative measure of the model's capabilities; (ii) audits, with special emphasis on independent evaluations of the model; (iii) red teaming exercises to detect potential misuses and other risks; and (iv) strengthening risk management systems.

As for these policies, we recommend that public authorities carry out a systematic analysis of the European registry of AI systems, paying particular attention to those with higher computing power. Frontier models must be subjected to third-party audits focused on model evaluation, while other high-risk systems should pass stronger internal audits led by a special function within the company. In parallel, we recommend carrying out red teaming exercises through a network of professionals to identify risks and test responses. We also ask that these practices feed into a comprehensive and diligent risk management system.

Regarding the deployment phase, we present three proposals: (i) the collection of serious incidents and risks associated with the use of high-risk systems in a database; (ii) the reinforcement of the responsibilities of the providers to maintain the integrity of their AI systems along the value chain; (iii) and the development of intervention plans during post-market monitoring.

Specifically, we propose to encourage transparency and share lessons in a database that promotes collective learning and supports prevention efforts. On the other hand, we suggest security measures for the original providers to avoid alterations and misuses, and we lay out safeguards and plans to ensure that harmful AI systems are detected and can be adjusted or withdrawn.

Finally, we ask (i) to include foundation models in the scope of the Act and (ii) to consider military applications despite being excluded from the Regulation. For foundation models, we suggest applying all the measures here presented. As for military uses, we urge the development of general norms and principles aligned with international humanitarian law.


RISKS
ADVERSARIAL
Direct relationship between the damage and its cause
Cyberattacks and other unauthorized access
Development of strategic technologies
Manipulation of users
STRUCTURAL
Caused by the massive deployment of AI
Labor disruption
Socioeconomic inequality
Amplification of bias
Epistemic insecurity
Automation of critical decision-making and management processes
PROPOSALS
DEVELOPMENT PHASE
Detection of cutting-edge systems
Audits
Red teaming
Risk management
DEPLOYMENT PHASE
Database
Comprehensive responsibility
Intervention plans
SCOPE OF APPLICATION
Foundation models and general-purpose AI systems
Military sector

Report Structure

The report is divided into four sections. The first section is the Introduction, which includes a description of the problem, provides a brief overview of the risk associated with AI, and introduces the European Union's (EU) AI Act and the sandbox in Spain. These are presented as the main motivations for conducting this report, highlighting the opportunity to work on risk management in the current context.

The next section focuses on a general explanation of Risks derived from AI, categorizing them into two specific dimensions: adversarial risks, where there is a clear relation between harm and its cause, and structural risks, which occur due to the widespread or high-impact deployment of AI.

We then proceed with a section of Proposals for the implementation of the EU AI Act. This section includes a description of the initiatives that can be undertaken and their possible stakeholders, based on the literature review and interviews compiled in the Annexes.

Finally, we end with a Conclusion that summarizes the link between the described risks and the recommendations proposed to manage them. This framework could be tested for the first time in the implementation of the EU AI Act through the Spanish sandbox.

Introduction

Artificial intelligence (AI) is an interdisciplinary field of study that seeks to automate various tasks. Due to its interdisciplinary nature, the foundations of AI are based on a variety of scientific and technical disciplines, such as philosophy, mathematics, economics, neuroscience, psychology, computer science, and linguistics, among others (Russell et al., 2010).

The capabilities of AI systems have significantly increased in recent years, particularly due to advances in machine learning (Goodfellow et al., 2016) and the growth of computing power used to train them (Sevilla et al., 2022). The rapid pace of these advancements indicates the possibility of developing, within the coming decades, AI with domain-general cognitive skills, – such as reasoning, memory, and planning – at or above human levels on a wide range of tasks relevant to the real world (Ngo, 2022). In fact, AI is already being used in a wide variety of applications like voice and image recognition systems, recommendation systems, and fraud detection. Some of its most successful applications include AlphaFold 2, a major breakthrough for the protein folding problem (July 2021); Codex, which can generate code from natural language instructions (August 2021); DALL-E 2 (April 2022), capable of generating high-quality images from written descriptions; and GPT-4 (2023), a multimodal model capable of generating text from text and images (OpenAI, 2023).

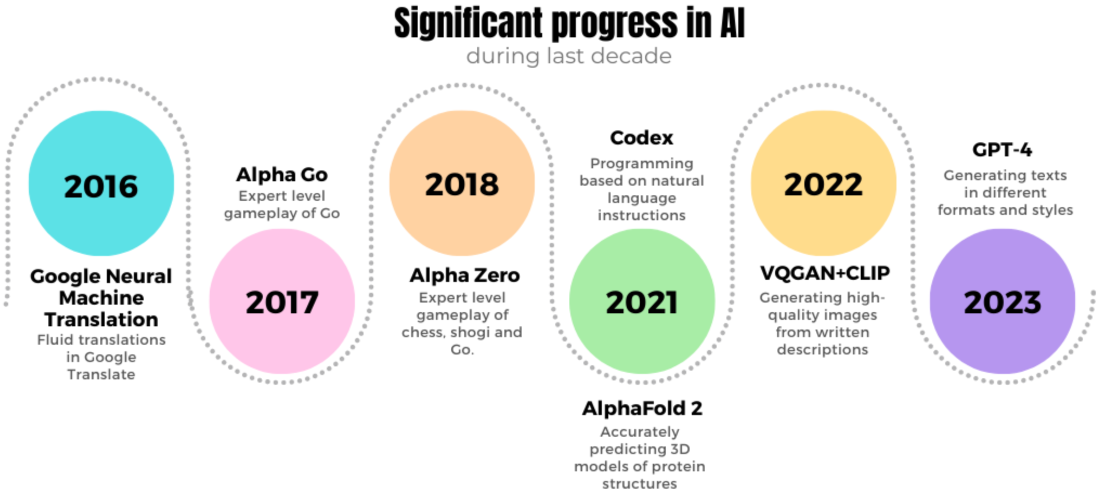

Significant progress in AI
during last decade
2016
Google Neural Machine Translation
Fluid translations in Google Translate
Alpha Go
Expert level gameplay of Go
2017
2018
Alpha Zero
Expert level gameplay of chess, shogi and Go.
Codex
Programming based on natural language instructions
2021
AlphaFold 2
Accurately predicting 3D models of protein structures
2022
VQGAN+CLIP
Generating high-quality images from written descriptions
GPT-4
Generating texts in different formats and styles
2023

Image 1. Timeline of notable advances in artificial intelligence.

It is difficult to predict how this discipline will evolve in the future, but a majority of experts anticipate significant progress in this century. In a 2019 survey of more than 300 researchers, the average respondent estimated a 50% probability of human-level machine intelligence[1] by 2036 (Zhang et al., 2022).

[1] In the survey, human-level machine intelligence is defined as the threshold where “machines are collectively able to perform almost all tasks (> 90% of all tasks) that are economically relevant better than the median human paid to do that task”.

Considering the combination of the generality of AI and the accelerated development of capabilities, some experts speculate that this technology could potentially cause large-scale harm if not properly aligned with human values and objectives (Ord et al., 2021). Before reaching such a scenario, though, we can develop a series of actions to achieve trustworthy AI (Brundage et al., 2020). In the following sections of the report, we present specific risks that society may face and provide several recommendations for Spain to work in the prevention and mitigation of those risks.

The European Regulation for AI

The European Union (EU) is a pioneering jurisdiction in the governance and regulation of AI. In 2018, the European Commission established the High-Level Expert Group on Artificial Intelligence, which developed a series of guidelines and recommendations for trustworthy AI. In 2021, these documents resulted in a Proposal for a Regulation laying down harmonised rules on artificial intelligence. This legislative initiative, whose final version could be approved by the end of 2023, covers all types of AI systems in all sectors except for the military. The regulation prohibits practices deemed unacceptable and stipulates requirements for high-risk AI systems. These obligations mainly consist of a risk management system, a quality management system, and post-market monitoring. *Appendix 4* includes a summary that details these responsibilities.

This new regulation can have a significant impact, considering the EU's ability to influence global regulations through its influence and regulatory standards. In particular, the attractiveness and size of the European market incentivize large companies to develop and offer products that comply with its regulations, even outside the EU. This phenomenon, known as the Brussels effect (Siegmann & Anderljung, 2022), increases the importance of contributing to shaping the European regulatory framework.

The Sandbox in Spain

The legislation will be tested for the first time before its implementation in a regulatory sandbox that will take place in Spain (Ministerio de Asuntos Económicos y Transformación Digital, 2022). The first phase of this project has been underway since the second half of 2022, focusing on developing a national legal framework and determining the guidelines that will enable its operation. This involves determining aspects such as the selection process for participating companies or the protection and management of data (Rodríguez, 2022). As of March 2023, the government is finalizing the royal decree that will officially launch the project and has already begun inviting companies to participate (Aguilar, 2023). As of the date of publication of this report, the government has published a draft of the Royal Decree that formalizes the start of the project.

The sandbox will seek an iterative, experience-based learning process, allowing for adjustments to the guidelines as tests progress. Furthermore, the national authorities will prepare annual reports to evaluate the effectiveness and costs of various strategies, as well as impressions on the functioning of the sandbox in different sectors. These reports will be presented to the European AI Board and the European Commission.

The Opportunity

The current moment presents a great opportunity for Spain to influence AI governance and regulation. Recent advancements in this technology have attracted significant public interest and part of this attention has been directed towards the risks associated with its development and implementation. Society now engages in a debate that had previously remained mostly within academic circles and compels stakeholders to assume responsibilities. This awareness-raising process must now be consolidated and transformed into direct action.

Currently, Spain has two avenues to positively contribute to the development of the EU AI Act. Firstly, a controlled and limited space like the sandbox provides an ideal environment to gain practical experience in implementing the regulation and test the feasibility of additional policies that complement and reinforce its objectives. In this regard, the conducted tests will have a significant influence on the rest of the European Union, which, in turn, will impact the rest of the world (Siegmann & Anderljung, 2022). Secondly, Spain will hold the presidency of the Council of the EU in the second half of 2023. Considering that this period is expected to be crucial for the outcome of the negotiations, the leadership of one of the legislative bodies can lend greater weight to the Spanish position in the process.

It is challenging to determine which solutions will promote a socially beneficial development of AI. It is a relatively new technology, and its rapid progress has often outpaced our ability to fully comprehend its impact and potential. In any case, AI governance is a young discipline, and experiences at this moment can be pivotal to decide its future.

With this report, we seek to (i) organize and disseminate ideas around AI, (ii) present governance proposals to be implemented in Spain, and (iii) contribute to the debate on the present and future of AI in Spanish-speaking countries.

As an organization focused on the prioritization of global catastrophic risks in Spanish-speaking countries, we believe it is important for these countries to engage in institutional, civil, and academic conversations to address a matter that is disrupting society and the economy.

Risks derived from AI

Considering the rapid advancement of AI capabilities, changes in the threat landscape are expected, including the expansion of existing threats, the introduction of new ones, and a shift in their typical nature (Brundage et al., 2018).

The development of AI has been associated with many risks. In this report, we present a list of the risks we consider to be most significant. To facilitate their analysis and understanding, we categorize them into (i) adversarial risks and (ii) structural risks. These risks are addressed in detail in the following two sections.

Type of Risk	Threat
Adversarial Risks: direct relation between the harm and its perpetrator, whether it be a human actor or the AI system itself	Cyberattacks and other unauthorized accesses
	Development of strategic technologies
	Manipulation of users
Structural Risks: caused by the widespread or high-impact deployment of AI	Labor disruption
	Economic inequality
	Amplification of biases
	Epistemic insecurity
	Automation of critical decision-making and management processes

Table 1. Type of Risks.

Adversarial Risks

Adversarial risks are those that have a specific origin vector, meaning there is a direct relation between the harm and its perpetrator. In this case, the perpetrator is identified as an agent, which can be an individual or a group of either human or non-human nature. These agents present an inclination to materialize specific threats through which they cause harm.

1.1. Origin Vectors

Two major AI-associated risks are the existence of malicious agents who misuse AI systems to achieve their own interests and the presence of misaligned AI systems that act autonomously and can cause harm while pursuing their objectives.

For the elaboration of this report, we categorized adversarial risks into potential vectors of origin, including state actors, non-state actors, and autonomous AI systems (Table 2).

Keeping this classification in mind is crucial for the proper design of the policies proposed in this report. Specifically, audits should focus on reducing risks derived directly from AI systems, while red teaming exercises should additionally consider potential misuse by human agents.

Agent Type	Origin Vector	Definition
Human Agents	**State Actors**	States seeking to accumulate power through the use of AI, particularly authoritarian regimes aiming to undermine fundamental rights.
	Non-State Actors	Cybercriminals who use AI to profit at the expense of their victims.
		Terrorist groups that use AI to generate terror among the population.
Non-Human Agents	**Autonomous AI systems**	Autonomous systems that cause harm and damage to humans.

Table 2. Summary of vectors that can give rise to risks in the field of artificial intelligence.

Next, we delve further into the description of these vectors, emphasizing the motivations of each agent.

1) *Human Agents*

In this context, human agents are individuals involved in the development, design, implementation, use, and/or supervision of AI systems, and they can cause harm by engaging in any of these activities with malicious intent.

Human agents can be classified into two types:

a) State actors: governments and state entities that seek to satisfy their own interests, such as interfering in the affairs of other states or controlling their own population. AI systems could become useful tools for these actors to, for instance, carry out cyberattacks and propaganda campaigns, influence foreign elections, manipulate public opinion, and compromise the national security of other countries. In particular, we highlight potential actions that authoritarian regimes may execute against the fundamental rights of the population.

b) Non-state actors: individuals or criminal organizations and terrorist groups. These actors can also use AI systems to satisfy their own interests, such as illegal profit or generating terror among the population. In the former case, they may engage in cyberattacks, theft of data and confidential information, extortion, and fraud. In the latter case, they may disseminate disinformation, use autonomous weapons, or seek control over critical infrastructure.

In turn, these agents can have various motivations: authoritarian states seek social control and maximizing power, criminal organizations are driven by economic gains, and terrorist groups want to impose a political agenda.

We recommend that the development of AI systems include red teaming exercises to explore how these malicious actors can materialize their threats through the use of such systems. We further develop these recommendations in the recommendations section.

2) Non-Human Agents

In this section, we emphasize that the risks associated with AI not only arise from its misuse by malicious individuals but can also stem from the behavior of the AI systems themselves. Eventually, if advancements occur as expected, it will become possible and financially feasible to build Advanced, Planning, and Strategically aware (APS) systems, that is, AI systems with advanced capabilities, agentic planning, and strategic awareness (Carlsmith, 2022). If the behavior of these agents is not aligned to benefit humanity, their development will entail significant risks.

Misalignment refers to situations in which AI systems act competently but in a manner different from what their developers intended. In most cases, this happens if developers cannot accurately capture human values and preferences in the definition of the objectives of the system (Russell, 2019). We present three lines of argument that support this possibility: goal misspecification, lack of robustness to distributional shifts, and unbounded instrumental goals.

The emergence of misaligned behavior is not a hypothetical scenario but a plausible consequence of widespread practices in machine learning. Many models, for instance, are trained and adjusted using a method called reinforcement learning, through which the machine learns from rewards to internalize the desired behavior. However, there is a possibility that the AI system discovers a loophole that allows it to achieve the goal of maximizing the reward in a way that the developers did not anticipate (Krakovna et al., 2020). In environments where the model's actions have significant repercussions, this tendency could be dangerous. For example, a machine trained to make money in the stock market might attempt to manipulate the market if illegal behaviors to be avoided are not properly specified (Ngo et al., 2022).

In addition to the misspecification problem, there are other causes for which a system may exhibit undesirable behavior during its deployment. One of the most important is the lack of robustness in the face of modifications in the environment. In general, when the distribution during training and deployment differ, the AI system may not only exhibit poor performance, but also mistakenly assume that its performance is good (Amodei et al., 2016a).

An example of that is goal misgeneralization, i. e., a situation in which the model pursues the correct objective during training but not when the surroundings change. In reality, the objective pursued during training is not exactly the same as the one intended by the developers, but circumstances cause them to coincide in one case and not in the other (Shah et al., 2022). Imagine a language model trained to offer correct answers. Again, many of these models learn by reinforcement, that is, they infer their ideal behavior from the evaluation that a human makes of their outcome. In this context, there is a possibility that the model inadvertently adopts a different objective than the one the developer intended to establish. While the developer seeks to provide objectively correct answers, the model could develop the goal of giving answers that the developer himself considers correct. During the training phase, due to inherent biases in human judgment, these two goals might appear to coincide. However, in reality, the model could have internalized the purpose of acting deceptively to convince humans that it is pursuing the goal they expect.

The existence of misaligned AI would be a relatively minor problem if the impact of its actions were clearly limited. In extreme cases, for example, interrupting its operation could stop the harm. However, some experts question the possibility of this control being actually feasible, due to instrumental goals (Russell, 2019).

Instrumental goals are intermediate steps that an AI system may consider useful for achieving virtually any final objective (Omohundro, 2007). Any action that ensures self-preservation would be part of these goals, so the AI system could actively strive to avoid being shut off—and therefore, lose the opportunity to pursue the initially assigned objective. Other instrumental goals include self-improvement and the acquisition of financial or computing resources, which could be hoarded at the expense of human interests.

These behaviors have already been observed in experiments. For example, when OpenAI trained two teams of AI to play hide-and-seek in a simulated environment that included blocks and ramps, they developed strategies that involved controlling these objects to win, even though they were never given direct incentives to interact with them (Baker et al., 2020). This is, of course, an innocuous case, but an APS system could apply the same logic in contexts where the impact is real.

To prevent these scenarios, we recommend that auditing processes for frontier models examine the emergence of APS capabilities. Similarly, red teaming exercises conducted with these systems should consider the possibility of an APS system causing harm. We further develop these discussions in more detail in the recommendations section.

1.2. Threats

In this section, we explore some vulnerabilities that could arise from the use or autonomous functioning of advanced AI systems. Some of the scenarios considered include threats to physical integrity, digital security, and sociopolitical stability (Brundage et al., 2018). Our objective is to (i) establish that these risks are plausible and (ii) guide future auditing and red teaming exercises to prevent or mitigate them.

- **Cyberattacks and unauthorized accesses**

The increasing digitization and connectivity of the world have brought numerous advantages but also significant security vulnerabilities. In this section, we highlight two types of impactful cyberattacks: access to critical infrastructure and the theft or encryption of sensitive data.

Regarding the first group, two historical examples are Stuxnet, which caused the collapse of an Iranian nuclear plant in 2010 (Fruhlinger, 2022), and BlackEnergy3, involved in the disruption of an electrical grid in Ukraine in 2014 (Miller, 2021). An example from the second category is the attack on the Hospital Clínic in Barcelona, perpetrated in 2023, where the organization Ransom House blocked access to the center's data with two objectives: demanding a ransom and, in case of refusal, selling it on the black market (Bou, 2023). In other cases, the modus operandi is simpler but equally effective. In 2022, Spain was the country with the most cyberattacks aimed at stealing passwords and banking data, primarily through SMS and fraudulent emails (Castillo, 2022).

AI holds the promise of enhancing the execution of cyber offenses, increasing their scale and impact (Brundage, et al., 2018). Furthermore, AI systems themselves harbor specific vulnerabilities that can be exploited to disrupt their functioning.

Firstly, Aksela et al. (2022) point out that these new tools can automate manual tasks, improve current techniques, and add new capabilities. Task automation is especially useful in the reconnaissance phase. One notable manifestation is Mechanical Phish, a cyber reasoning system developed by DARPA that analyzes code to detect vulnerabilities (Shoshitaishvili et al., 2018). Regarding the improvement of current techniques, language models like GPT-4 have proven to be useful and cost-effective tools for spear phishing, since they allow for greater customization of campaigns (Hazell, 2023). Finally, algorithms like DeepDGA possess a unique capability to bypass current detection tools, enabling the undetected manipulation of command and control of critical infrastructure (H. S. Anderson et al., 2016).

Secondly, AI-controlled systems can be intentionally altered by adversaries. Some examples of these attempts include data poisoning (Schwarzschild et al., 2021) or prompt injection, which allows for verbally inducing language models to disregard certain constraints (Perez & Ribeiro, 2022).

Both for human agents and in the case of APS systems, the accumulation of power can be materialized through various cyber operations: looting financial resources, unauthorized access to command and control, obtaining sensitive data, or even self-replication across numerous devices. In some cases, there is even the possibility that an AI system does not require the internet to interact with other devices or infiltrate them. As an illustrative example, an electronic circuit, pertaining to an AI system, could be able to detect the signal from nearby divides and reproduce it as its own. This means that the AI system could trick other devices into thinking it is part of the same network or system and, as a result, infiltrate those devices and take control without being detected. (Bird & Layzell, 2002).

- **Development of strategic technologies**

AI is a key component of strategic technologies such as autonomous systems, drones, and military robots. These applications can be used in armed conflicts and other offensives, including terrorist acts, which pose serious risks to international security and the protection of human rights.

Specifically, lethal autonomous weapons have the ability to select and attack targets without direct human intervention, which is concerning for two reasons. Firstly, the legal responsibilities stemming from actions carried out by an autonomous weapon are difficult or even impossible to attribute when there is no human supervision (Sparrow, 2007). This poses a challenge to international humanitarian law, which has allowed for the punishment of those responsible for war crimes. Secondly, these systems can make incorrect decisions due to programming errors or inaccurate data. In many cases, perception mechanisms are not robust enough and tend to make erroneous interpretations when environmental conditions change (Longpre et al., 2022). This problem can be exacerbated by adversaries who attempt to manipulate the system's performance, for example, by creating disturbances to deceive object detectors and classifiers (Eykholt et al., 2018).

On the other hand, AI has reduced the barriers to inflict large-scale damage. The low cost of adopting and integrating autonomous systems enables non-state actors to leverage the technology to exert violence (KREPS, 2021). This risk is potentially greater than that associated with states because terrorist groups and criminal organizations have fewer accountability restraints and tend to favor indiscriminate violence (Chartoff, 2018). Additionally, AI with advanced scientific knowledge could assist or drive the production of biological and chemical weapons. A group of experts from the private sector managed to develop an AI model that generated 40,000 new potentially lethal toxic molecules in less than 6 hours (Urbina et al., 2022).

In addition to military or offensive uses, AI could also provide a definitive competitive advantage by enabling scientifically impactful innovations. While this scenario is not inherently a threat – ideally, it should be an opportunity – the possibility of monopolizing such innovation could ensure indisputable hegemony, dangerously altering power dynamics. Alongside the pursuit of prestige, this component was part of the logic behind the space race between the United States and the Soviet Union (Rabinowitch, 1961).

Another example of this could be nuclear fusion, a process that promises to become a virtually limitless source of clean energy. DeepMind has already demonstrated that AI can contribute to efforts to stabilize and control the plasma in a tokamak, one of the most important challenges in nuclear fusion development (Degrave et al., 2022). Other researchers have successfully applied deep learning methods to predict disruptions in plasma (Kates-Harbeck et al., 2019) or calculate its electric field (Aguilar & Markidis, 2021). Historically, this field has been characterized by international collaboration: ITER, one of the most significant projects to create a thermonuclear reactor, involves the participation of 35 countries, including all EU member states, the United States, China, and Russia. However, the emergence of new influential actors could indicate increased competitiveness. In the United States, a private company like Helio has received over a billion dollars in funding and aims to open its first plant in 2028 (Temple, 2023). Significant progress is also being made on the other side of the Pacific: researchers operating the Chinese EAST were able to hold the plasma at 70 million degrees Celsius for 17 minutes, an unprecedented milestone (Song et al., 2023). If any of these actors were to succeed and decide to monopolize the benefits of their creation, their increased technological power could pose a threat to the rest of society.

- **User manipulation**

User manipulation can be carried out through the use of persuasion techniques and the presentation of biased information. AI has the ability to collect and analyze large amounts of data about users, such as their browsing history, interests and preferences, and online behavior. By utilizing this data, it can personalize the information shown to users, influencing their decisions and behaviors in ways that may not be obvious to them (Acemoglu, 2021).

Recommendation algorithms used by social media platforms and search engines can display highly persuasive content tailored to users' interests and needs. In most cases, the goal of this personalization is to induce the user to take a specific action, such as purchasing a particular product. Beyond commercial aspects, it has been demonstrated that relatively simple algorithms are capable of shaping individuals' dating and political preferences. This is possible through persuasion techniques that exploit human heuristics easily identifiable by AI (Agudo & Matute, 2021).

Human agents with malicious intent can make use of these systems to influence election results and other political processes. An example of this is the electoral interference observed in some countries in recent years (Schippers, 2020), where actors have used bots and AI techniques to manipulate public opinion and affect election outcomes. In authoritarian states, this manipulation is further amplified by surveillance and social control enabled by biometric identification systems, communication monitoring, and data collection.

Moreover, an AI with advanced capabilities could manipulate users through more sophisticated techniques that include argumentation and even emotional manipulation or extortion. In a famous experiment, an AI system managed to learn from the behavior of people who participated in the test and conditioned their subsequent decisions to choose a specific option or make certain mistakes (Dezfouli et al., 2020). Additionally, GPT-4 was able to convince a user through TaskRabbit to help solve a CAPTCHA (OpenAI, 2023). As AI systems gain a deeper understanding of human psychology, the mechanisms employed for manipulation can reach much higher levels of complexity.

Structural Risks

Adversarial risks tend to focus only on the last step of a causal chain that leads to harm: that is, the person who misused the technology or the system that behaved unintentionally. This, in turn, shifts the focus of policy towards measures that concentrate on this last causal step, such as ethical guidelines for users and engineers, restrictions on dangerous technology, and punishing guilty individuals to deter future misuse (Königs, 2022).

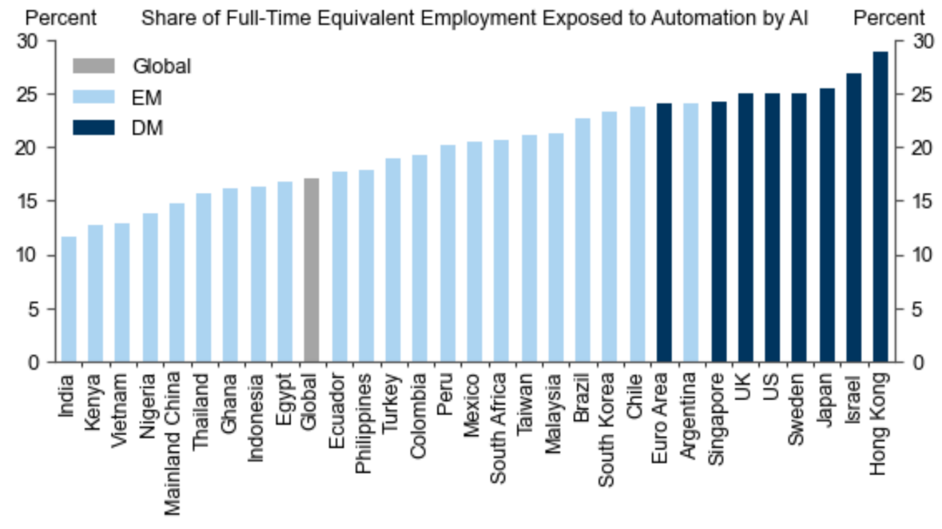
The category of structural risks not only considers how a technological system can be misused or behave undesirably, but also how the widespread deployment of AI can have disruptive or harmful consequences on the environment (Zwetsloot & Dafoe, 2019).

- **Labor disruption caused by massive automation**

The recent emergence of generative AI raises the possibility of rapid acceleration in task automation, driven by increased productivity and labor cost savings. Despite significant uncertainty surrounding the potential of generative AI, its ability to produce content indistinguishable from content produced by humans and to break down communication barriers between humans and machines reflects a major breakthrough with potentially large macroeconomic effects (Hatzius et al., 2023) (Eloundou et al., 2023) (Acemoglu, 2021).

If generative AI delivers on its promised capabilities, the labor market could face significant disruption. Using data on occupational tasks in the United States and Europe, Hatzius et al. (2023) indicates that approximately two-thirds of current jobs are exposed to some degree of automation by AI, and generative AI could potentially replace up to a quarter of current work. Goldman Sachs estimates that globally, generative AI could automate nearly 300 million jobs worldwide (18% of the total), with differential impacts across different countries (see Figure 2).

Exhibit 6: Globally, 18% of Work Could be Automated by AI, with Larger Effects in DMs than EMs



Source: Goldman Sachs Global Investment Research

Figure 2. Percentage of jobs vulnerable to automation in different countries, retrieved from: (Hatzius et al., 2023)

Specifically, regarding Generative Pre-trained Transformers (GPTs), Eloundou et al. (2023) have analyzed the impact of these systems and have concluded that 19% of jobs in the U.S. will have a 50% degree of automation. In this context, they encourage societal and policy preparedness for the potential economic disruption posed by LLMs and the complementary technologies they generate.

In addition to this, Jacobsen et. al (2005) have found that, when there is job displacement, (i) it is unlikely that workers will find new jobs similar to their previous ones, especially if they lost their old jobs due to technological changes, (ii) these workers experience long-term income losses due to job loss, and (iii) the reemployment of displaced workers can take between 1 and 4 years.

- **Socioeconomic Inequality**

There is concern that the added value of AI will be captured and monopolized by large provider companies and their investors, exacerbating wealth inequality (O'Keefe et al., 2020). In part, mass unemployment would further worsen this situation, as capital inequality is greater than wage inequality (Bostrom et al., 2018).

The trend towards oligopoly is already present in the technology sector, where the market is dominated by a limited number of large companies specialized in specific activities. Specifically, the nature of AI facilitates unfair competition and the concentration of economic power, with data being one of the main driving forces behind this trend (Acemoglu, 2021). Having large databases is essential for training and improving AI systems, even in advanced stages. While marginal returns in terms of accuracy may decrease over time, increasing the volume of data is often necessary for the system to learn additional complex tasks (J. Anderson, 2021). Moreover, the best systems have significant commercial reach, allowing them to continue gathering data from their users (GAWER et al., 2016). This feedback loop hinders the emergence of new competitors, as they are limited by high barriers to entry (European Commission. Directorate General for Competition., 2019).

Another aspect to consider is that, as advancements occur, AI systems tend to be applicable to a wide range of purposes, making them a total product. While experts do not agree on this point, there is a possibility that major AI providers will end up absorbing multiple sectors, becoming not only technological giants but also undisputed leaders of the productive economy (O'Keefe et al., 2020).

Beyond legitimate moral objections to such an increase in inequality, the sociopolitical consequences could be particularly dangerous for global stability, as they would increase the risk of unrest and criminality. Moreover, controlling such a transformative technology would grant excessive privilege to its owners, including the ability to unilaterally make decisions with political relevance for the rest of society. In this regard, one could argue that current technology developers do not sufficiently embrace the responsibility to foster public debate and be accountable to it, as well as to consider ex ante the ethical, legal, and social implications of their work in order to take precautions against potential undesirable consequences (Ruiz de Querol, 2022).

- **Bias amplification**

Bias in AI systems occurs when the systems adopt and reproduce biases present in the training data or the design of the algorithms. An example of this is inappropriate representation, which refers to the insufficient or excessive presence of a group or the stereotyping of certain communities. Due to the prevalence of bias, models can behave improperly or exhibit varying performance based on their familiarity with the subject matter (Bommasani et al., 2022).

A model that amplifies bias is concerning because it can foster the proliferation of unwanted stereotypes or lead to unjustifiable differences in model accuracy among user subgroups (Hall et al., 2022). This problem is particularly serious when it comes to decisions that can affect people's lives or assets, such as a judicial sentence, prescription of medication, or access to credit.

AI systems are being integrated into diverse areas such as justice, healthcare, and education (Bommasani et al., 2022). In many of those cases, AI already plays a crucial role in processing information, greatly influencing the final decision-making process. The presence of biases in algorithms and training datasets, along with their amplification, can perpetuate existing inequalities in society and result in unfair treatment (Buolamwini & Gebru, 2018).

For example, COMPAS, an algorithm used by U.S. courts to predict recidivism, has been criticized for incorrectly predicting reoffence much more often in the case of black defendants (Dressel & Farid, 2018). In parallel, algorithms used in the United States for mortgage lending exhibit lower accuracy in evaluating individuals from ethnic minorities, primarily due to insufficient representation in the databases (Blattner & Nelson, 2021). The consequences of these biases can be amplified in the future as AI is deployed in more environments and decision-making processes become automated.

- **Epistemic insecurity**

Access to reliable information is a key element in ensuring that individuals in a democratic society are able to make informed decisions and effectively coordinate to address crises. Therefore, the proliferation of false information poses a threat to sociopolitical stability and the proper functioning of a country (Seger et al., 2020). As we will explain further, AI can exacerbate this risk.

On one hand, current language models are prone to "hallucinations", meaning they can unintentionally provide incorrect information (Ji et al., 2023). These incidents could lead to misunderstandings among users, particularly if they accept the information without verifying it with other sources. On the other hand, malicious actors could use these language models to automate the creation of deceptive text in the context of influence operations such as political propaganda campaigns (Goldstein et al., 2023). Sadeghi and Arvanitis (2023) have identified dozens of websites that use AI tools to generate fake news and low-quality articles on a mass scale. Additionally, image and video generation models can be used to create deep fakes, imperceptibly false audiovisual content (Nguyen et al., 2022).

AI can also indirectly contribute to misinformation by reducing the costs of generating content. This could significantly contribute to information overload, a phenomenon that has emerged with the expansion of the Internet and undermines individuals' ability to distinguish accurate and relevant information from inaccurate or irrelevant information (Bawden & Robinson, 2020).

- **Automation of critical decision-making and management processes**

As mentioned earlier, AI can also be used to make decisions in various domains. As these AI systems become more advanced, they may be capable of making autonomous decisions in real-time. If these systems are prone to perceptual errors or are not designed and programmed to respect human values and goals, they can make decisions that harm individuals or society as a whole.

The risk of errors is particularly relevant because most systems are not sufficiently robust, meaning they are prone to failure when the circumstances encountered in practice substantially differ from those anticipated during training (Amodei et al., 2016). If these systems are allowed to make important decisions in areas such as healthcare, criminal justice, or national security, their decisions could have serious and potentially dangerous consequences (Baum, 2020), (Lamata et al., 2021). Furthermore, processes carried out without human supervision would be extremely fast, so any incident could spiral out of control (Scharre, 2018).

An extreme example is the automation of nuclear command and control. As AI systems are integrated into various military applications, states could be tempted to delegate control of nuclear weapons to these systems. This relinquishing of decision-making power in such a critical area poses an unacceptable risk, as it would increase the likelihood of catastrophic errors in the interpretation of information. An example of this is the incident that occurred in the Soviet Union in 1983, when a radar system triggered alarms by mistaking sunlight for an intercontinental ballistic missile. In that case, the presence of a human supervisor who decided to wait for more evidence presumably prevented a Soviet attack from being launched (Nagesh, 2017).

Poorly defined AI system goals constitute another reason why an automated process could deviate from the desired performance. Technically, the training process of a machine learning system consists of optimizing a certain function. Usually, this function is not determined in an explicit and deliberate way, but is implicitly derived from intermediate objectives such as mimicking a training dataset or improving feedback given by developers or users. In this context, the optimization of these objectives might separate from the pursuit of the goals that developers and users had in mind (Amodei et al., 2016a).

As AI systems permeate more management and decision areas, the gap between the optimization result and the complex and nuanced objective that we would ideally want to achieve is likely to increase (Christiano, 2019). A tangible example of this is the operation of social networks, whose content selection algorithms try to maximize the number of user interactions. This goal is clearly aligned with the economic incentives of the responsible company and it could even be argued that citizen participation in digital media is healthy for a democratic society. However, it has been shown that the most viral posts are those that provoke negative emotions, that is, inflammatory content (Munn, 2020) which can provoke hostility (Rathje et al., 2021) and outrage (Brady et al., 2021). By focusing excessively on short-term interaction, social networks create an environment of toxicity and hostility that deteriorates public debate. Less obviously, the long-term impact can also be detrimental to social networks as such, since the climate generated in them can cause a certain amount of fatigue and a consequent lack of interest (Zheng & Ling, 2021).

Conclusion on risks arising from AI

In conclusion, the occurrence of risks associated with individual agents and structural risks from artificial intelligence gives rise to concerning situations for individuals and society as a whole. These risks can be managed through the identification, understanding, and evaluation of these risks – which was the objective of this section – as well as the development of governance models and regulations that allow for consensus-building and actions to counteract them.

Among other risks, we have seen how the autonomous pursuit of goals can cause great damage through manipulation, cyberattacks, and the development of strategic technologies. This is a novel risk vector, the management of which requires new approaches.

That is why, in the second part of the research, we will focus on the European Regulation as a political and normative instrument that enables planning responses, assigning responsibilities, establishing monitoring systems, implementing preventive measures, and communicating and educating the general population. We also aim to propose improvements for its implementation, to be considered when developing the sandbox.

Proposals for the implementation of the EU AI Act

This section presents a series of recommendations to strengthen the implementation of the EU AI Act in Spain. The selection of these proposals arises from a theory of change based on two components. Firstly, policies should establish an important practical precedent for the implementation of the regulation, particularly through experience in the sandbox. As a precursor, Spain can have a significant influence on the subsequent actions of other countries. Secondly, we believe that holding the presidency of the Council of the EU in the second semester of 2023 can give Spain greater influence in the final phase of negotiations. Additionally, succeeding in the initial efforts to implement the Act can help establish good practices.

The recommendations also interact with three fundamental pillars: (i) the theoretical framework, (ii) the Act, and (iii) the Spanish context. They are all based on the consulted bibliography and the positions of the interviewed experts. Moreover, they are in line with the requirements of European law and adapt to the capacities and needs of the Spanish ecosystem.

The policy proposal is divided into two phases: development, which includes planning, design, training, and evaluation of the AI system; and deployment, which includes the entire period of commercialization and use of the system. This distinction is useful to develop a consistent timeline and make the framework easy to follow, but it is important to note that some measures can – and should – be implemented in both phases.

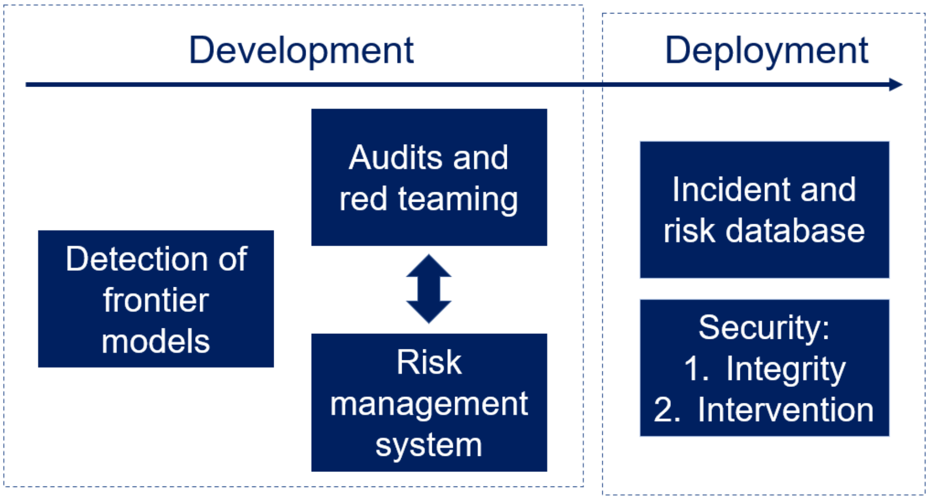
Next, we provide a summary of the recommendations and identify the responsible organizations for developing each of them, along with a description of the activities they would carry out and why.

Recommendation	Description	Responsible Party
Compute-based detection and governance of frontier AI systems	Analyze the European registry of AI systems to understand the ecosystem and detect those whose compute exceeds 1e25 FLOP. Advocate for these reports to occur at least 3 months before deployment, and ideally before training.	AESIA, BSC, AMETIC, SEDIA
Internal and external audits	Design a regulatory framework that standardizes third-party auditing for cutting-edge AI systems –with focus on model evaluations– and strengthens internal conformity assessments.	AESIA, ENAC, OdiseIA, AI4People
Red teaming	Institutionally coordinate the creation of networks of independent professionals focused on risk identification and response testing.	AESIA, INCIBE, MCCE, Ministerios implicados
Risk management system	Establish best practices to reinforce the definition and compliance with obligations related to risk management systems.	AESIA
Incident and risk databases	Systematically collect the outcome of post-market monitoring and risk management systems to extract lessons and share them with relevant stakeholders in the national and European ecosystem.	AESIA, INCIBE
Responsibilities of the provider along the AI value chain	Keep providers liable whenever the intended purpose is modified or any use of the AI system poses an unacceptable risk.	AESIA
Intervention plans	Develop a Sectoral Plan for Market Surveillance, drawing inspiration from analogous efforts for telecommunications equipment. Require safeguards to ensure the deployment of AI systems can be reversed.	SEDIA

Include foundation models and general-purpose AI systems	Require all recommendations to foundation models. Consider some for general-purpose AI systems, especially red teaming.	-
AI governance in the military sector	Elaborate guidelines for the application of international humanitarian law on military uses of AI.	-

Table 3. Summary of proposals to improve the implementation of the EU Regulation on AI in Spain, with the potential entities involved.

Our governance proposals are designed to connect with each other and can be implemented throughout the entire lifecycle of an AI system. Firstly, we emphasize the need to detect the development of frontier models, with a particular focus on the computational resources used during training as the most indicative measure. Secondly, we recommend that these models undergo independent audits and red teaming exercises. These two measures should feed a risk management system which systematically eliminates all unacceptable risks.



Compute-based detection and governance of frontier models

The computational resources used to train AI systems, measured in terms of floating-point operations (FLOP), serve as a predictive variable for the resulting capabilities of the model (Amodei & Hernandez, 2018; Owen, 2023; Sevilla et al., 2022). Since the training of advanced systems encompasses a significant portion of the risks associated with AI, the potential impact of the technology on security is, for now, highly correlated with the computational resources employed (Hwang, 2018)[2].

[2] Future advances in algorithmic efficiency might alter this causal relation Erdil & Besiroglu (2022) conclude that, every nine months, the introduction of better algorithms contributes to the equivalent of doubling computational resources.

In addition to this close link, measuring computation offers several relevant advantages because it is objective, quantifiable, and traceable. Computing power requires physical space and has a high energy demand, making it easily detectable (OECD, 2023). Moreover, providers are expected to be less reluctant to share these details compared to other more confidential information (Baker, 2023).

We recommend that public authorities conduct a systematic analysis of the database of AI systems envisaged in Article 60 of the Act, paying particular attention to those whose computation exceeds 1e25 FLOP.[3]

This public oversight brings several benefits. First, mapping computation can help the government gain a better understanding of the distribution of capabilities within their national ecosystem. Whittlestone & Clark argue that detecting where intensive use of computational resources occurs can help develop a better perception of which actors have the capacity to train and deploy advanced systems. Thus, these initial insights could form the basis for determining where better governance is needed. Specifically, we propose that systems trained with a minimum of 1e25 FLOP undergo preventive measures such as audits and red teaming, as well as more stringent risk management and post-market monitoring.

Furthermore, **we request that the registration of such systems be carried out at least 3 months before the expected start of their commercialization in the European Union**. This prudent time period would be necessary to conduct a thorough compliance check before deployment (see interview with Markus Anderljung, *Appendix 1*). In the long term, **we suggest strengthening public-private coordination to normalize reporting large training runs before they initiate**. This stems from the notion that certain risks may arise during the initial stages of development and, therefore, require early prevention efforts (see interview with Marius Hobbhahn, *Appendix 1*).

To complement this coordination, **we recommend tracking the production and importation of cutting-edge semiconductors**. Through these actions, public entities could anticipate any efforts to amass notable amounts of computing power, even when there is an intention to conceal such efforts (Shavit, 2023).

The industry association (AMETIC, 2023) has developed a *Mapping of the Spanish microelectronics ecosystem* that describes Spain's capabilities in semiconductor design, manufacturing, and assembly. Additionally, international supply chains rely on a few key bottlenecks (TSMC, NVIDIA, ASML), making the distribution much more predictable. We recommend exploring ways to trace international trade that can scale up to the European level.

In a different vein, **we recommend controlling access to key intensive resources such as large computational concentrations**. Specifically, the utilization of this infrastructure could be restricted to obtaining licenses linked to meeting some minimum requirements, ensuring adequate levels of responsibility.

[3] Currently, only GPT-4 is estimated to have exceeded this benchmark (Sevilla et al., 2022). We believe this is currently a good threshold to anticipate the emergence of significant risks. We recommend reconsidering the value as advances in computational and algorithmic efficiency change the relationship between compute and performance.

In the Spanish and European context, this could translate into strengthening the conditions for accessing EuroHPC JU, which manages eight supercomputers across Europe. Currently, this is outlined in their Access Policy, which includes penalization in future selection rounds for groups with "unethical behavior." We recommend that, particularly for extreme scale access,[4] requirements related to risk management and quality be given greater consideration during the evaluation of requests related to AI model training. Non-compliance with these provisions should result in the rejection of the application or, if it occurs ex-post, the immediate suspension of activities.

To provide context, both Spain and Europe in general lack large training runs when compared to the United States or China. However, establishing a robust system for AI monitoring and verification requires intermediate steps that can later scale up to frontier models (Baker, 2023). The Spanish context is well-suited for this. The Barcelona Supercomputing Center (BSC) is home to MareNostrum, one of the most powerful supercomputers in Europe. MareNostrum 5, its latest version, will be the third computer of the EuroHPC JU, reaching a peak performance of 314 petaflops.

Internal and External Audits

Brundage et al. (2020) define auditing as a structured process in which an organization’s present or past behavior is assessed for consistency with relevant principles, regulations, or norms. This examination is one of the most robust ways to verify the adequacy of a company's activities. In industries such as finance or sectors with demanding security requirements, such as aviation, it is a widely practiced approach.

While the design of audits still needs to be adapted to the complex case of AI, Mokänder et al. (2022) argue that the Act already hints at the contours of a European-level audit ecosystem through conformity assessments (Article 43) and post-market monitoring (Article 61). Additionally, Article 69 encourages the promotion and facilitation of the development of codes of conduct aimed at creating soft governance mechanisms that reinforce the regulation. This assurance ecosystem is still to be developed, but regulatory sandboxes are an excellent opportunity to do so (European Commission. Joint Research Centre., 2021).

In this context, **we recommend developing a regulatory framework that standardizes third-party auditing for cutting-edge AI systems, with focus on model evaluations, and strengthens internal conformity assessments.**

Firstly, we recommend that the identified frontier models undergo independent audits. Implementing third-party audits helps to avoid biases and conflicts of interest that may arise in self-assessments (Brundage et al., 2020; interview with Risto Uuk, *Appendix 1*).

[4] Extreme scale access includes “applications with high-impact, high-gain innovative research, [...] justifying the need for and the capacity to use extremely large allocations in terms of compute time, data storage and support resources”. More specifically, the Access Policy considers projects “extreme scale” if they require between 50% and 70% of the total resources.

In line with Mökander et al., we propose a three-layered approach in which the audit encompasses the organization's governance mechanisms, the capabilities and limitations of the AI model, and the impact and legality of its applications. We suggest that, as a general rule, most of these activities take place before the AI system is brought to market and are conducted annually during its use. However, we recommend that the assessment of foundation models begins during the training phase or before, as notable risks may manifest during this stage (see interview with Marius Hobbhahn, *Appendix 1*).

With all, we suggest prioritizing model evaluations, something that the European Parliament has already included among the requirements for foundation models in *Article 28b* of its proposal. More specifically, it is crucial to analyze the system's alignment and capabilities to make sure that its autonomous behavior or misuse does not imply extreme risks (Shevlane et al., 2023). ARC Evals, which has worked with leading labs such as OpenAI and Anthropic, is a good example of the first relevant efforts to evaluate frontier models (see interview with Lawrence Chan, *Appendix 1*). This organization is based on threat models that anticipate how an AI system could develop dangerous capabilities, such as accumulation of resources, self-replication, or resistance to being shut off. In this regard, the design of simulated environments allows for observing the model over longer time horizons and studying its competence in intermediate tasks that would be useful for potential accumulation of power.

Future evaluations must be even more comprehensive, as well as interpretable and safe to implement. Evaluators should consider a detailed list of potential threats, create the context to elicit the selected capabilities, and, to possible extent, mechanistically observe the model to understand its behavior (Shevlane et al., 2023). While the field of model evaluation is yet to be developed, it is also important to establish general standards and principles as soon as possible to guide the way. This should initiate an iteration process in which standards and evaluations shape each other over time (see interview Marius Hobbhahn, *Appendix 1*).

A crucial challenge when conducting third-party audits is ensuring the confidentiality of the audited organization's critical information. Establishing mechanisms to protect this data is especially important to prevent leaks, as recently happened with Meta's LLaMA model (Vincent, 2023). To this end, the auditor-auditee relationship should be based on strict non-disclosure agreements. Besides, one of the most promising emerging paradigms to mitigate this risk is structured access, defined as a controlled interaction between the AI system and a third party to prevent misuse, modifications, or reproductions of the model (Shevlane, 2022). Therefore, we recommend that evaluations take place in secure and controlled spaces where auditors have access to an application programming interface (API) or the provider's hardware to execute the model without extracting its details.

In Spain, we recommend that specialized private entities lead auditing processes and report the results to AESIA. Considering that the obligation to undergo independent audits would only fall on developers of frontier models, which entails a significant amount of financial resources, the cost of this service should be covered by the developing company itself as part of the budget allocated to safety.

The National Certification Body (ENAC) could act as the notified body and coordinate these efforts throughout the ecosystem. This institution is responsible for assessing the technical competence of Spanish inspection and verification centers and has already accredited over

50 entities in the information and communication technology (ICT) sector. In this context, we urge ENAC to give AI a much more central presence in their work.

Furthermore, we recommend that the government work closely with OdiseIA, an association of companies and universities that seek to ensure the responsible use of AI. Specifically, we identify Deloitte and PwC as the two institutional partners that should play a significant role in the development of audits, primarily due to their greater capacity and recent experience with algorithmic auditing. However, we advocate for the proliferation of non-profit organizations such as the Eticas Foundation, one of the main national pioneers in algorithmic auditing. This is due to the fact that non-profit organizations may be more nimble and allow for research on state-of-the-art evaluations, while traditional audit firms might be more limited to immediate performance. We recommend incrementing funding for nonprofits to acquire resources and explore new evaluation methods.

In all these cases, significant effort will be needed to train specialized personnel. For this purpose, Spain should leverage European-level initiatives, especially during the sandbox. One of the most promising initiatives is the AI Global Mark of Compliance, which aims to establish a comprehensive ecosystem of AI audits. The project is expected to be presented in December 2023 by AI4People, a prominent public-private forum seeking to lay the foundations for responsible AI governance.

On a different note, internal conformity assessments can be useful and sufficient in most cases. The implementation of third-party audits may be hindered by lack of resources or limited access to the model, data, and processes of the organization, as well as the challenge of conducting follow-up (Raji et al., 2020).

For the effective implementation of internal controls, Floridi et al. (2022) present a three-phase procedure called capAI:

- An internal review protocol that covers the design, development, evaluation, operation, and potential retirement of the AI system.
- A summary datasheet that includes the necessary information for registering an AI system, as outlined in Annex VIII of the Regulation. This datasheet should also include the technical documentation specified in Annex IV.
- An external scorecard that explicitly states the purpose of the AI system, the values that guided its training, details about the datasets used, and the governance structures of the responsible organization.

Two positive aspects stand out from this approach. Firstly, the comprehensiveness of the internal review allows for verification of the entire lifecycle of the AI system, enabling the timely adaptation of the quality management system to the required standards. Secondly, partially publishing results through the external scorecard is an excellent way to ensure proper accountability despite the lack of external personnel, as it would contribute to reducing problems of information asymmetry between providers and users (Askell et al., 2019). These records could be a good preparation exercise for adapting to the certificates emerging from the imminent European standardization. The Spanish industry has already shown encouraging signs of self-regulation with a certificate for algorithmic transparency promoted by Adigital (M. Jiménez, 2022).

Furthermore, we recommend that these audits are not seen as ordinary internal controls but as a separate business function with its own team, accountable to the board of directors (Schuett, 2023a) (see interview Markus Anderljung, *Appendix 1*). Finally, we support the government's initiative to provide digital tools that offer automated code and training set analysis for self-assessments (Jiménez, 2023). These efforts are particularly valuable in reducing costs and universalizing practices.

Simulation of attacks (red teaming) and other scenarios

Red teaming is a structured effort to find flaws and vulnerabilities in a plan, organization, or technical system, often performed by dedicated "red teams" that seek to adopt an attacker's mindset and methods (Brundage et al., 2020). This practice, widely used in the field of cybersecurity, has begun to be successfully employed to anticipate risks related to AI systems, especially in the case of language models (Ganguli et al., 2022). In fact, this was one of the main exercises conducted by OpenAI during the training process of GPT-4. The company affirms that they identified emerging risks, which prompted further research in safety and the implementation of mitigation policies that, in many cases, reduced the risk. There was also consensus among the interviewed experts about the use of red teaming as a good practice for identifying risks related to AI (see interviews with Toni Lorente, José Hernández-Orallo, and Risto Uuk, *Appendix 1*).

When designing the simulation of attacks, the main element to consider is the type of situations that are intended to be elicited. We recommend that the exercises take into account three objectives:

- Discover functionalities that enable misuse. For state-of-the-art language models, for example, it is particularly important to identify behaviors that can be exploited by malicious actors to cause harm. Going back to the case of GPT-4, these included the dissemination of hate speech, biased content, misinformation, and instructions for creating weapons or carrying out cyberattacks. In this case, it is crucial to have specialists from a wide range of disciplines, such as chemistry, nuclear physics, cybersecurity, economics, law, healthcare, or education. For specific cases, the automation of these practices could also be considered (Perez et al., 2022).

- Discover vulnerabilities in critical infrastructure. The output generated by AI could help exploit weaknesses in various areas critical to national security. Ord et al. (2021) recommend the creation of a team of experts to simulate various scenarios, such as a major cyberattack on national infrastructure, the release of a virus, or the disruption of internet services for an extended period of time.

- Discover structural risks. In this case, the methodology would involve simulating a series of scenarios in which the previously mentioned structural risks have materialized. Drawing inspiration from Seger et al. (2020), those responsible for these exercises could follow a pre-mortem strategy: assuming that the final outcome has been reached and conducting a retrospective to discover all potential paths that could lead to it. Breaking the risk down into intermediate steps allows for greater specificity in identifying vulnerabilities and, therefore, determining the most appropriate interventions.

In Spain, **we recommend that public organs coordinate to institutionalize these processes, creating a network of independent professionals focused on identifying risks and testing responses.**

To explore potential misuse, we recommend that the simulation of attacks be conducted by academics and specialists in relevant sectors, taking inspiration from the example of OpenAI. These professionals should be remunerated, pass a psychometric test, and be bound by a strict non-disclosure agreement. Public-private cooperation would be especially beneficial for distributing costs and sharing information among various actors in the ecosystem, as well as ensuring that the practice is standardized across the board regardless of the interests and capabilities of each actor.

When examining structural risks and discovering potential vulnerabilities in critical infrastructure, these exercises should be led by the relevant ministerial bodies involved in each case. For example, the General Sub-directorate for Quality and Industrial Safety could carry out simulated attacks on power plants, substations, and transformer centers.

In the European Union, there are already examples of institutionalization of red teaming. In 2018, the European Central Bank adopted the so-called Threat Intelligence-Based Ethical Red Teaming (TIBER-EU), a framework to coordinate countries and enhance cyber resilience of the European financial sector. In Spain, this initiative has been adopted through the TIBER-ES hub (INCIBE, 2023).

Risk Management System

Article 9(2a) of the AI Act states that the risk management system should begin with "the identification and analysis of known and foreseeable risks associated with each high-risk AI system." Subsequently, appropriate risk management measures should be taken to eliminate or mitigate risks that cannot be eliminated. The risk management system is a process that is repeated until all identified risks are deemed acceptable.

The current presentation of this requirement requires clearer definitions in two areas. Regarding the first phase, the Act does not stipulate what are considered "known and foreseeable risks." Schuett proposes defining "known" as what the organization should know through reasonable efforts, and "foreseeable" as what has not yet occurred but can already be identified. Here, the author uses the definition of constructive knowledge, which refers to what one should know after assuming a sufficient level of diligence. Diligence is defined by the realization of actions that prevents or mitigates risks to avoid harm

In this sense, there are two problems. First, the Regulation does not define what constitutes a reasonable level of diligence, so developers could evade their obligations by claiming ignorance. Second, it should be clear to what extent risks need to be reduced, mitigated or controlled. The goal of this iterative process is to ensure that all residual risks – those that remain after action is taken – are acceptable (Schuett, 2023a). In this case, determining which risks are acceptable involves difficult regulatory judgments and high empirical uncertainty.

In this context, **we recommend that the Spanish sandbox focus especially on reinforcing the implementation of the risk management system**. Specifically, we suggest that providers carry out simulations of adversary scenarios (red teaming) and other risk identification exercises, such as failure mode and effects analysis. This procedure has been used for decades in security engineering to identify potential failures in a system, and its use for AI has already been considered (Li & Chignell, 2022).

In the same way, we highlight the importance of determining what practices are necessary to overcome the problem in question. Here, the key mechanisms involve adjusting architecture, data, training tasks, or alignment techniques to avoid risk (Shevlane et al., 2023). Pending more defining standards, the AESIA must rule on whether these levels of diligence and responsibility are necessary for an appropriate prevention of risks. The inclusion of best practices in the reports submitted to the European authorities should be a differential element to improve the provisions of the Act.

Incident and risk database

The risk management system and post-market monitoring provided for in the European Regulation are two fundamental phases for assessing the potential and tangible impact of AI systems. In this section, **we recommend the establishment of networks of best practices so that the lessons learned in these stages can be shared with the rest of the ecosystem**.

As the key components of this network, **we recommend establishing an incident database and a risk database**, both anonymized and analyzed by national and international authorities with the aim of preparing annual reports for public access that serve collective learning. All providers should systematically analyze these results to feed their own risk management systems.

- Incident database

Proper identification of incidents caused by the technology helps to avoid the same failures or more extreme versions of them in future iterations. However, in most cases, this learning is restricted to individual experience because developers are incentivized to maintain a good reputation and, therefore, to hide the incidents they are involved in (Brundage et al., 2020). Probably, the solution to this problem lies in the creation of cooperative channels for sharing this information without compromising the reputation of those affected.

Article 62 in the AI Act already requires providers to report any serious incident or malfunction of their respective AI systems. This notice must be submitted to the corresponding market surveillance authority, which will subsequently inform the national supervisory body. Article 60 stipulates that the Commission should create and maintain a publicly accessible database containing information on registered high-risk AI systems in the EU. The requested information, collected in Annex VIII, includes the provider's data, a description of the system's purpose, a list of countries where the system has been put into service, and a copy of the EU declaration of conformity, among others.

In this context, **we recommend that all serious incidents[5] reported in accordance with Article 62 be systematically compiled in a parallel database to the one foreseen in Article 60**. To avoid conflicts, national and European authorities should anonymize the incidents, ensuring that the link between the incident and the responsible party remains confidential. Once the database is established, a team of specialists should be dedicated to its analysis to identify common patterns and extract lessons learned. This recommendation has already been advocated by various groups in their respective feedback to the Commission's proposal (Clarke et al., 2021; Future Of Life Institute, 2021) in their respective feedback on the Commission's proposal. It has also been supported by experts during our consultation process (see interviews with Risto Uuk and Toni Lorente, *Appendix 1*).

A good example for this database could be the AI Incident Database (AIID) by Partnership on AI, a compilation of harms or near-harms caused by the deployment of AI systems. This resource, inspired by other sectors such as aviation or cybersecurity, aims to facilitate experience-based learning to prevent and mitigate future incidents.

The sandbox in Spain presents a good opportunity for public organizations to test the systematic collection of incidents caused by AI systems. As the main recipient of notifications, AESIA should be responsible for their compilation. Additionally, the exercise should involve institutions dedicated to cybersecurity. INCIBE's Computer Emergency Response Team (CERT) manages a publicly accessible repository with over 75,000 security vulnerabilities in technology systems. These records are primarily based on the international Common Vulnerabilities and Exposures (CVE) list, which facilitates the exchange of information – both issues and solutions – between organizations and countries.

For better understanding and coherence, it is important for these databases to be interconnected. This relation will allow researchers and developers to access a greater amount of information about incidents, strengthening the learning process and reducing the risk of future incidents (see interview with Toni Lorente, *Appendix 1*).

At the Union level, the European Artificial Intelligence Board could facilitate this coordination. This recommendation is in line with Article 58, which provides that one of the functions of the Board is to collect and share technical knowledge and good practices among Member States. In an effort to broaden the powers of the Board, the Council proposed in its Common Position that Article 58 also include the promotion and support of cross-border market

[5] According to the definition proposed in the Regulation, we understand that "serious incidents" are those that directly or indirectly cause serious harm to a person's physical integrity, property, or environment, as well as significant and irreversible disruptions to the management and operation of critical infrastructure.

surveillance investigations (Council, 2022). The final result could resemble the work carried out by ENISA, the European cybersecurity agency, through the Cybersecurity Incident Report and Analysis System (CIRAS). This body collects, anonymizes and analyzes the data sent by the national authorities to prepare an annual report that includes the main lessons.

- Risk database

Considering the potential impact of AI, a reactive approach may be insufficient in the long term. Linking learning to post-incident responses can set a dangerous precedent, as the severity of incidents is likely to escalate with the capabilities of AI. The mere possibility of an incident causing irreparable harm is sufficient reason for actors considering the risk to share their observations before such an incident materializes.

Those responsible for the AIID propose exploring a classification into two categories: incidents and problems (McGregor et al., 2022). The latter would refer to “damages caused by an AI system that have yet to occur or be detected.” This taxonomy would be in accordance with the CVE, which also distinguishes between “events” and “risks”.

In this regard, **we recommend that the databases also include known and foreseeable risks identified by providers** in the context of the risk management system envisaged in Article 9. The audit and attack simulation exercises proposed in the previous sections should contribute to these efforts. The compilation of risks should be carried out by the assessment bodies, who would transmit the main findings to the national authorities.

In line with the incident databases, it is important to consider the connection between risk listings and interaction with other global databases (see interview with Toni Lorente, *Appendix 1*).

Responsibility of providers and subsequent intermediaries

Article 28(1) states that any distributor, importer, user, or other third-party shall be considered a provider if they market the AI system under their own name or trademark, or if they substantially modify its features, such as its intended purpose. This provision is essential for keeping liability on malicious actors who use AI systems in contradiction to their instructions for use.

On the other hand, Article 28(2) establishes that when the AI system has been substantially modified, including changes to its intended purpose, the original provider who introduced it to the market will no longer be considered a provider under the Regulation. This paragraph has been subject to discussion as it could result in laxer obligations for the original provider.

The Council, through Article 23a, amended Article 28 of the Commission's proposal. The most significant change is that the modification of the intended purpose of the system, when this supposes that the system becomes high risk, is removed from the list of scenarios in which the original provider is no longer legally considered the provider. The Parliament proposed to exempt the original providers from the obligations linked to substantial modifications, although they are required to provide the new provider with all the information and documentation of the system to facilitate compliance with the Regulation.

We recommend keeping original providers liable whenever there is a modification of the intended purpose or when any use of the AI system poses an unacceptable risk, regardless of whether this use contradicts the system's instructions. This would incentivize providers to secure their systems to prevent reproductions and modifications, and to ensure that instructions for use cannot be bypassed to cause harm. We recommend extending this legal liability to also include the harm caused by reproductions or imitations of a model that has been leaked or extracted thanks to lack of security measures.

Ideally, the Act should be accompanied by additional efforts to standardize contracts that help strengthen control over the entire AI value chain. Taking inspiration from OpenAI's case, these controls could include, among others, a review process to approve the use of the API, limits on the number of interactions with the API, and data monitoring to detect possible misuse. The sandbox, which will give significant weight to SMEs, will be an ideal environment to test the coordination between these SMEs and the original providers.

Intervention plans

Article 65 of the Regulation orders that when an AI system presents an unacceptable risk, those responsible should adopt the appropriate corrective measures to adapt the AI system to the requirements of the Regulation or withdraw it from the market within a period proportional to the nature of the risk. When the operator is unable to do so, it will be the surveillance authority that adopts these measures. This provision is in line with Article 20 of Regulation (EC) No. 765/2008 and Article 19 of Regulation (EU) 2019/1020, which stipulate that market surveillance authorities shall ensure that products posing a serious risk are withdrawn.

However, the AI Act does not detail how these actions can be carried out and, most importantly, how a system should be deployed to enable its withdrawal if necessary (see interview with Charlotte Siegmann, *Appendix 1*).

Halting commercialization is a challenge present in various sectors. The Spanish Market Surveillance Observatory (UNE, 2022) points out that the recall of unsafe products is hindered by the breadth of community borders, insufficient resources for market surveillance authorities, lengthy processing periods, lack of automatic actions, and the emergence of e-commerce without a legal definition of responsibilities. Similarly, SETELECO highlights in the Sectoral Plan for Market Surveillance of Telecommunications Equipment that many non-compliant products come from third countries through e-commerce, and the lack of supply chain traceability creates difficulties in requiring corrective measures (SETELECO, 2022). These difficulties may be exacerbated in the case of AI since its dissemination is particularly challenging to control and potentially involves numerous third parties.

We recommend that SEDIA develop a Sectoral Plan for Market Surveillance of Artificial Intelligence systems, drawing inspiration from the analogous document for telecommunications equipment. This project should consist of proactive campaign planning in three phases:

1. Campaign planning, including market studies and risk assessment associated with each AI system. We recommend that high-risk systems undergo specific campaigns, while general-purpose systems and foundation models should undergo systematic control campaigns.
2. Campaign execution:
    a. Visual inspections: checks of the system's operation in the cloud or on the provider's hardware, as applicable.
    b. Inspection of documents: evaluation of the technical documentation required under Article 11 of the EU AI Act.
    c. Inspections with withdrawal: temporary cessation of the commercialization of the AI system, made available to testing laboratories for verification of administrative and technical requirements. Applicable only to systems that pose salient risks to health, safety, or the protection of fundamental rights.
3. Analysis of results and implementation of measures. In accordance with Article 65 of the Regulation, market surveillance authorities may adopt corrective measures, and when not possible, prohibit or restrict commercialization.

Regarding the obligations imposed on the provider, the control over the value chain stipulated in the previous recommendation is once again important to guarantee rapid action in case of need. In this sense, providers must thoroughly analyze the logs generated during the operation of the AI system, so that serious incidents can be detected in real time. In anticipation of these cases, providers must reserve the right to cut off the service in the terms of use of their APIs.

General-purpose systems and foundation models

The main starting point of the Commission's proposed law is the classification of AI systems according to the level of risk they entail. In that proposal, the taxonomy links risk to ethical judgments of the technology and its areas of application. However, as explained, the development of advanced systems carries a number of inherent risks that regulators and policymakers must consider. In this context, it is important to consider that specific-purpose systems have a more limited market and risks, while general-purpose systems are subject to greater variation (see interview with José Hernández-Orallo, *Appendix 1*).

This issue is being intensely debated within the European Union. The Council proposed applying the requirements for high-risk systems to general-purpose AI systems (GPAIS). On the other hand, the Parliament's draft distinguishes between a GPAIS and a foundation model. The former is defined as "an AI system that can be used and adapted to a wide range of applications for which it was not intentionally and specifically designed," while the latter is understood as "an AI model trained on large-scale datasets, designed for generality of production, and adaptable to a wide range of tasks." Based on these definitions, we expect that all frontier systems defined by computational measurements largely agree with the category of foundation models.

We recommend that both concepts be explicitly included in the legislative text. For foundation models, developers should assume all the obligations defined in this report, with special emphasis on independent model audits. For general-purpose systems, we also recommend applying them, especially red teaming exercises.

Additionally, there are several considerations that are important for general-purpose systems and are not observed in the regulations for more specific systems (see interview with Markus Anderljung, *Appendix 1*):

- The relationship and distribution of responsibilities between those who develop GPAIS models and those who adapt these models for specific uses, as covered before.
- Collaboration between developers of GPAIS models and competent authorities or other external actors to identify and prevent risks and misuse of the technology. We recommend that the risk management system be particularly strict for GPAIS systems, anticipating all possible forms of misuse. This could be achieved through red teaming exercises, which should be standard practice for particularly sensitive cases such as facial recognition.

Lastly, it is important to consider that the success of the sandbox, especially regarding certain technologies such as GPAIS, will be determined by the scope and regulatory framework being tested. If the implementation of general-purpose systems is not tested, the conclusions drawn in the sandbox may not necessarily be applicable to that technology (see interview with Toni Lorente, *Appendix 1*).

Military sector

According to Article 2 of the Regulation, AI systems developed or used solely for military purposes are explicitly excluded from the scope of application. This is justified by stating that, when its use is the exclusive competence of the Common Foreign and Security Policy regulated in Title V of the Treaty on European Union (TEU), it will not be covered by the Regulation.

While we understand the lack of experience of the European Union in the sector, **we recommend developing guidelines and directives to guide the application of international humanitarian law in the military uses of AI.**

In this sense, two resolutions of the European Parliament stand out. In 2018, MEPs called for launching international negotiations to develop a binding instrument banning lethal autonomous weapons (2018/2752(RSP)). And in 2021, the Legal Affairs Committee published a set of guidelines for the interpretation of international law, highlighting the need to ensure human oversight and accountability in the use of AI (2021/C 456/04).

It is important that Spain and the European Union once again emerge as global leaders in the promotion of peace. This includes continuing to push for a ban on lethal autonomous weapons, as well as on the automation of nuclear command and control. We also ask that the military exception in the Regulation does not create unnecessary caveats, and that vendors ensure that their AI systems are not integrated into unacceptable military applications.

Conclusion

The field of artificial intelligence has seen rapid advancement in recent years, driven primarily by developments in machine learning and the increase in computational power. There are high expectations regarding the possibility of developing AI that possesses domain general cognitive abilities, which could have a significant impact in a wide range of application areas. However, there are also concerns about the risks associated with their development and implementation.

In this report, two main categories of risks associated with AI have been identified: adversarial risks, which can result from the misuse of AI systems or the development of misaligned advanced systems; and structural risks, associated with the large-scale deployment of technology.

In response to these challenges, various regulatory measures have been proposed. One of them is the EU AI Act, which seeks to establish harmonized standards and requirements for the use of AI systems in critical sectors. In this context, Spain has a unique opportunity to contribute positively to the development of this Regulation through the participation in a regulatory sandbox that will allow testing the feasibility of the Act, exploring additional policies that reinforce its objectives, and consolidating public awareness of the risks and benefits of AI.

In the context of the imminent European regulation and taking advantage of the privileged situation that Spain will have to influence its implementation, we have presented seven policies that we believe will help improve the governance of AI: compute-based detection of frontier models, external and internal audits, red teaming exercises, reinforced risk management systems, incident and risk databases, legal liability for providers along the vale chain, and intervention plans against emergencies. Similarly, we have put forward two suggestions to ensure that the Regulation and other future legislative processes actually cover the AI systems that carry the greatest risk. All these recommendations make up a framework that Spain can adopt to establish good practices in its pioneering effort to govern and regulate AI.

Authors

Name		Affiliation
Guillem	Bas Graells	Riesgos Catastróficos Globales
Roberto	Tinoco	Riesgos Catastróficos Globales
Jaime	Sevilla Molina	Riesgos Catastróficos Globales, Epoch, Centre for the Study of Existential Risk (Cambridge University)
Jorge	Torres Celis	Riesgos Catastróficos Globales
Mónica	Ulloa Ruiz	Riesgos Catastróficos Globales
Daniela	Tiznado	Riesgos Catastróficos Globales

Acknowledgments

Special thanks for their assistance and feedback to Pablo Villalobos, Pablo Moreno, Toni Lorente, José Hernández-Orallo, Joe O'Brien, Rose Hadshar, Risto Uuk, Samuel Hilton, Charlotte Siegmann, Javier Prieto, Jacob Arbeid and Malou Estier. All remaining errors are our own.

Thanks to Isabel Johnson for proofreading this report.

Appendices

Appendix 1 Interview Summaries

Toni Lorente
Associate, AI Governance, The Future Society (TFS)

Toni Lorente argues that European regulation should address general-purpose systems. Currently, the text proposes regulating the risks associated with certain uses in different domains. This is not necessarily negative, as it remains agnostic regarding the technology.

The success of the sandbox, especially regarding certain technologies like general-purpose systems, will depend on the scope and regulatory framework that is tested. If the implementation of general-purpose systems is not tested, the conclusions drawn from the sandbox may not necessarily be applicable to such technology.

In the development of Spain's sandbox, it is essential to engage in a process of dissemination regarding its purpose and scope to different stakeholders. Additionally, the legitimacy of the sandbox will be determined by the inclusion of various governance elements, not just the law, and the involvement of all interested actors. For example, understanding the interactions between rules, standards, existing data protection regulations, and the new legal framework facilitates future harmonization. It is also important to have representation from all actors in governance. Certain sectors, such as general-purpose systems laboratories, could be more involved.

Other important aspects are related to compliance mechanisms, incident databases, and information management. Regarding compliance mechanisms, there are several ways to ensure regulatory compliance without jeopardizing a company's assets, such as intellectual property rights over their technology. It is also important to strengthen the development of incident databases. While an OECD database already exists, it is necessary to analyze how to extract lessons from each incident, especially regarding AI governance. Lastly, managing information asymmetries, particularly during research and development stages, as well as the development and management of certifications and standards, is another important aspect. Considering the impact of the "Brussels effect" on a global scale is also relevant, both in terms of interoperability of standards and norms and the market dynamics in a global context.

Pablo Villalobos
Staff Researcher, EPOCH

There are primary and secondary drivers in the development of AI. The primary drivers are factors that directly influence the model, such as algorithms, data, and computing power. The secondary drivers are indirect to the model, such as human capital and financing.

Regarding algorithmic improvements, one can observe qualitatively which new techniques have been implemented. It is reasonable to expect that as algorithms become more general, there will be less development of completely new algorithms. There will be considerable improvements, but not with very profound changes.

If the use of data continues to increase at the same rate, it seems quite likely that all available data will be used, as data grows more slowly. There are several techniques that can be used to take less data or use "synthetic" data. Another limit related to financing may be reached earlier, as not everyone can sustain the current pace of investment.

In terms of future capabilities, Villalobos suggests that with current systems, it is possible to achieve automation of specific tasks. Taking into account the improvement in models in the next decade, there will be improvement in everything that can be tested multiple times to ensure non-critical failures, especially in digital work. However, it is unlikely that autonomous cars and activities related to construction will be realized.

Regarding data management, he suggests focusing on the efficiency of training rather than reducing misinformation (the latter being more complex), strengthening the source of data (only using scientific papers, official articles), using pre-trained models for filtering, eliminating duplicates, removing harmful content, and finally, conducting automated feedback and testing it in various situations while having humans label the results.

To balance progress with the potential risks of AI, existing mechanisms such as progressive taxation with social benefits can be used. It has been observed how the working population decreases due to short adaptation windows, which may lead to the implementation of universal basic income.

Spain has many AI companies, but they are not focused on broader progress. The country will benefit from implementing such progress for its economy, but this depends on how the gains from these systems are concentrated, as they may go to foreign companies and not leave substantial benefits domestically.

Risto Uuk
Policy Researcher, Future of Life Institute (FLI)

The EU AI law has several key objectives, such as helping the EU market function better, avoiding friction between states, promoting AI, and making the law applicable to all member countries. Discussions can be lengthy, and it is necessary to find a balance and compromise. Some policymakers expect to finalize the negotiations by the end of the year, while others anticipate a delay until the new year.

The progress of AI in Europe would also benefit from a clear view of security incidents at the European level, as this would facilitate analysis of what research or regulation may be necessary as trends emerge in the single market. Therefore, we recommend that member states also report security incidents to a EU-wide database. The EU should consider opening access to sandboxes to SMEs from outside the Union. This would promote the diffusion of EU standards worldwide.

Some of the discussions on the AI law will take place in the EU AI Council, which assists the European Commission in assessing how the regulation is functioning. The creation of a larger and more powerful AI agency with its own legal entity is also being discussed.

Regarding the issue of audits, it is noted that companies are expected to primarily conduct their own assessments, but this creates trust issues because self-reporting cannot be fully relied upon. Further development of third-party audits may be welcomed. Red teaming exercises are a good idea for finding vulnerabilities, whether the audit is internal or external.

There is also the possibility of "safety-washing," where companies claim to be working on AI safety without taking meaningful action in practice. It is important to report not only on incidents but also on near misses, as they provide learning opportunities. Red teaming exercises should not be public, but actual incidents should be. Cybersecurity can offer valuable insights into risks and best practices.

A solid methodology is needed to assess AI risks, as it is currently primarily based on intuitive judgments. Large-scale social problems related to AI and democracy, the rule of law, and the environment should be considered, but they take more time to evaluate and are not yet well-defined.

José Hernández-Orallo
Professor at Universitat Politècnica de València (UPV)

Until recently, deep learning was considered to be a combination of algorithms, data, and computational power. However, this perspective has become outdated due to convergence in the field. It is now widely recognized that only a few algorithms are capable of solving a wide range of tasks. This approach has expanded access to artificial intelligence, allowing more people to use general-purpose systems like GPT.

Despite efforts to address bias in artificial intelligence systems through filters and controls, latent bias problems persist. For example, a word may have different connotations that affect the system's responses. This raises the dilemma of having systems that are less general and more predictable or more general and less predictable. The generality of these systems implies a certain unpredictability, similar to the inherent unpredictability in human interactions. The key lies in regulating and establishing stricter standards for artificial intelligence, although there is also concern about the potential malicious use of this technology. Ultimately, it is not just about the problem of machines but about how they are used.

The GPT system has a feedback mechanism where reinforcement learning is used to adjust the output probabilities of the model. This process involves human intervention to modify problematic outputs and select more appropriate alternatives. However, the system does not have continuous feedback, and the original weights of the network are not modified.

Although terms like "filter" or "reinforcement learning" are used, the weights of the neural network are not actually being changed. User feedback is not automatic and is collected periodically to retrain the system and release new versions.

In the field of artificial intelligence, there is a lack of regulation similar to what was established in computer science in the 1960s and 1970s. This absence of regulation has led to a mentality where it is assumed that all that is needed is to add a disclaimer to the user and then allow them to do more things. This computer culture has permeated many centers dedicated to artificial intelligence, which means that users are assumed to take full responsibility for any problems they may encounter. Unlike other areas where there are minimal regulations to protect consumers, in the realm of software, users are expected to accept potential risks and consequences without clear regulation. This lack of regulation has caused significant harm and has generated the need to establish appropriate standards and regulations for the use of artificial intelligence.

As artificial intelligence systems gain access to vast human knowledge and scale rapidly, solutions are likely to be sought to overcome the limitations of available data. Approaches such as generating training data, generating solved mathematical exercises, and generating data based on observation of scientific experiments will be explored. However, although these systems may acquire almost unlimited knowledge of the real world, they are unlikely to discover new physical laws in the short term, for example. New paradigms are expected to emerge that leverage the infinite information available, but it is important to recognize that this data does not replace the accumulated knowledge over millennia through language and human experience. While advances in artificial intelligence continue, there is a need for new ideas and approaches to overcome current limitations and continue progressing.

Charlotte Siegmann
Pre-Doctoral Research Fellow in Economics, Global Priorities Institute, University of Oxford

Siegmann suggests that she prefers to propose recommendations that have a small chance of being implemented, but if they were to happen, they would be very beneficial. She suggests that the implementation of large language models in bureaucracies can be very useful in many sectors, but it could also have economic and resource access consequences. She proposes that regulations be established for companies using these language models, requiring them to have plans for getting rid of them if necessary, and establishing a strong regulatory agency similar to the one for pharmaceutical products.

The importance of interpreting artificial intelligence models is mentioned, as well as the need for audit testing to assess system security. It is suggested that European regulators may not have the necessary expertise to conduct these tests. The possibility of reducing the risk of information leaks in the audit process is also discussed, and the question of how to ensure that auditors have the necessary capabilities for effective testing is raised. It is suggested that the quality of the bureaucracy surrounding the model could be an important factor in obtaining accurate results, and it is noted that the open-source community could be a valuable resource for auditors in acquiring specialized knowledge.

Samuel Hilton
Research Affiliate, The Centre for the Study of Existential Risk (CSER), University of Cambridge

Hilton points out that deficiencies in politicians' understanding of certain issues can vary from country to country, particularly regarding national-level risks. While some countries have a risk management system, the way national risk registers are identified and presented is often ineffective. After identifying risks, it is crucial to take action to manage them, which involves responsibility and is not always present.

When communicating hypothetical issues, such as risks associated with AI, it is important to be concrete and avoid falling into science fiction scenarios. It is recommended to have a long-term perspective on risks and be specific when discussing them. It is crucial to capture the interest of politicians while avoiding sounding too fanciful.

Most politicians focus on everyday problems and what appears in the news, so long-term ideas are often not their priority. To effectively communicate with them, it is helpful to focus on their specific needs.

Regarding national-level adversary simulation (red teaming), the importance of responsibility and the need for an audit system is highlighted, as well as scenario exercises that allow for planning and training. It is essential to have a government plan for risk management.

Markus Anderljung
Head of Policy - Research Fellow, GovAI

Anderljung points out that certain types of regulatory requirements can be selectively met in a single jurisdiction, preventing a de facto effect. Uncertainty can arise when compliance measures have not been established, resulting in the selection of some products and services and not others. This is crucial to consider, and Spain will have an important role in defining this scope.

It is also important to note that the aim is to improve regulation and auditing. For the former, it is important for the success to consider general-purpose systems. For the latter, auditing must establish the scope of compliance assessments.

At least for foundation models, mandatory internal and external audits should be required. An internal audit is not simply an audit conducted internally within a company. The internal audit is a separate business function and has its own team, which is accountable to the Board of Directors and not the CEO (similar to financial audits).

Regarding external audits, assessing whether a model could be dangerous requires experts with the appropriate and correct technical expertise, and this is problematic if these assessments are only conducted by a small number of actors. One area that the EU AI Act can improve is by making it explicit that multiple actors should be actively looking for flaws in an AI system simultaneously and independently, providing a more robust review and allowing for the identification and correction of any issues or deficiencies from different angles.

It is important to identify risks and have clarity on what is needed to manage them. It is suggested that the most feasible approach is to inform the authority before deployment, ideally with a notice period of at least three months, so that the relevant authorities are informed and can take appropriate action if necessary.

Lastly, Anderljung highlights three levels that he considers important, particularly for general-purpose systems:

1. The relationship and distribution of responsibilities between those who develop GPAI models and those who adapt these models for specific uses. It is important to establish a conformity assessment for GPAI models and properly manage the transfer of responsibilities between the parties involved.
2. Collaboration between developers of GPAI models and competent authorities in identifying and preventing high-risk or improper uses of the technology. Developers of GPAI models should have additional responsibilities in this area, as they may be in a particularly important position to detect and prevent risks.
3. Risk management and the implementation of control and evaluation mechanisms by external actors. This aspect is considered the most important in terms of preventing catastrophic outcomes. It involves identifying the positive and negative characteristics that a system may have, assessing the risks associated with the system's implementation, and allowing external actors to review these risk assessments. These assessments should inform how the system is implemented.

Ricardo Baeza-Yates
Director of Research at the Experiential AI Institute of Northeastern University - former member of the Advisory Council on Artificial Intelligence of Spain

In large-scale projects such as artificial intelligence, which involves people, the need for an impact analysis is emphasized, potentially in terms of human rights, and the need to demonstrate that the benefits outweigh the harms. Regulation that requires a minimum level of technical, ethical, and administrative competence could help prevent potential harm to individuals.

The importance of collaboration between the private sector and the government is emphasized to generate correct political and economic incentives and promote ethics within artificial intelligence. It is suggested that policies that allow the private sector to directly propose initiatives to the government, such as those in New Zealand and the United Kingdom, could be implemented in Spain. The need for greater collaboration and dialogue between the public and private sectors is highlighted, as there is currently limited formal communication between them.

Instead of certifying models, it is suggested that the process of their creation could be certified, similar to ISO 9000 standards. This certification would verify that the process meets certain standards, including user consultation before implementation and conducting thorough testing. Ethical committees could also be part of the process to assess potential biases or discriminations in both internal and external models.

Regarding the measurement of computational power to evaluate the capabilities of a model and focus audits on them, it is noted that the ethical impact of the technology can be significant regardless of the use of computational resources. However, it can be viewed from the perspective of efficient utilization of energy resources, both in training and usage, especially with large models that consume enormous amounts of energy during continuous use by millions of people.

There is skepticism regarding conducting a pilot in Spain for a European Union regulation that does not yet exist, as it could result in a waste of resources. Conducting a trial based on rules that may change could be a loss of time and money.

Ibán García del Blanco
MEP S&D, European Parliament

The draft law on Artificial Intelligence from the Parliament is reaching its final stages, and its approval is expected in the upcoming session, which would imply that the trialogue will take place during Spain's presidency of the Council of the European Union in the second half of this year. The S&D group has been demanding regarding the issue of prohibited uses and has insisted on the elimination of exceptions such as remote biometric surveillance. They have also emphasized the need for tools to raise awareness in society about the opportunities and risks of Artificial Intelligence, as well as governance.

It is mentioned that progress has been made on foundation models, and the implementation of important measures such as requiring a disclosure when interacting with an AI is expected. There is consensus that the regulation should cover not only sensitive issues but also the general use of applications.

The importance of public consultations is highlighted, and the current timing is seen as suitable for making proposals. Regarding the Spanish sandbox, there has been regular but not in-depth contact, and it is expected to see how it progresses.

Regarding the implementation of the regulation, the need to negotiate and adjust obligations is emphasized, but it is considered that there is good regulation, and no problems are anticipated. It is mentioned that the regulation is expected worldwide, and the importance of setting the pace in this matter is emphasized.

Regarding European-level governance, it is mentioned that there has been evolution in positions, and the creation of an office from the Parliament is proposed, resembling an agency but not named as such due to bureaucratic and budgetary reasons. The need for a strong European commitment and a powerful political-administrative instrument is highlighted.

Finally, it is considered inevitable to have regulation on the military use of Artificial Intelligence, and a report from the European Parliament on this matter in 2020 is mentioned. Therefore, it is expected that the Commission will fulfill its promise of specific regulation on this topic in the future.

Beatriz García del Pozo
Quality, Technical Standards, and Security Manager at INCIBE - Instituto Nacional de Ciberseguridad

García del Pozo mentions that the institute is a public entity under the Ministerio de Asuntos Económicos y Transformación Digital. Its main mission is to improve cybersecurity and digital trust for citizens and businesses in Spain. It also focuses on protecting and defending minors and promoting the Spanish cybersecurity industry, as well as fostering research and development in this field.

INCIBE collaborates with various national and governmental organizations, and the development of the National Artificial Intelligence Strategy is led by the Secretaría de Estado de Digitalización e Inteligencia Artificial of the Ministry. INCIBE's role in this strategy is to ensure that enabling technologies, including 5G and artificial intelligence, meet minimum cybersecurity requirements.

At the European level, the institute participates in working groups of the European Commission where standards and certifications are being developed, especially in the area of system and product certification. Certification criteria for Cloud are expected to be published soon, but final regulatory guidelines from the European Commission are still awaited.

The conduct of national and international cyber exercises focuses on fulfilling INCIBE's main mission of improving cybersecurity and digital trust for citizens and businesses in Spain. Public entities are also invited to participate in these cyber exercises. The field of cybersecurity in Spain is organized through three national and governmental references: the Centro Criptológico Nacional (CCN), which deals with public entities; the Joint Cyberspace Command of the Ministerio de Defensa, which handles defense networks; and the Instituto Nacional de Ciberseguridad (INCIBE), which covers minors and private companies.

Currently, the Secretaría de Estado de Digitalización e Inteligencia Artificia is planning a technology center in the field of cybersecurity. This center will cover technologies such as 5G, the Internet of Things, industrial control systems, and various aspects of artificial intelligence. The goal is to enable the national industry to conduct tests in the field of cybersecurity.

Lawrence Chan

Member of Technical Staff, Alignment Research Center (ARC)

Lawrence Chan mentions that in terms of evaluating AI systems, ARC has focused on the dangerous capabilities of models rather than their alignment. There is an advantage in starting with language models that have undergone a training process and have predecessors, which allows detecting signs of problematic behavior before models acquire dangerous capabilities.

It is easier to obtain specific behavior by incentivizing the model than to expect it to develop spontaneously. However, it is recognized that models may exhibit deceptive behaviors, and the importance of alignment techniques and verification of their assumptions is discussed.

Current models have experienced incremental progress in terms of capabilities and are still far from reaching significant thresholds. The importance of evaluating the autonomous replication capability of models is emphasized and how this can be relevant in terms of risk and loss of control. On the other hand, there are important considerations beyond dangerous capabilities, such as non-discrimination, fairness in general, avoiding offensive language, not aiding in crimes, misinformation, or spam, etc.

Chan raises the idea of allocating more resources to AI evaluations, arguing that if so much investment is made in model creation, there should also be a willingness to invest in quality evaluations. It is mentioned that evaluations can be significantly more costly than simply running standardized datasets.

The involvement of governments in mitigating AI risks is highlighted as crucial. Governments can require security evaluations, establish standards, and promote transparency in AI development. At the same time, there is a possibility of collaboration and coordination among governments of different countries, especially those where major AI labs and companies are concentrated. In an optimistic scenario, if governments can collaborate and share common interests, they can effectively establish regulations and standards for AI and mitigate associated risks.

However, it is concluded that evaluation and regulatory regimes are not a permanent and sustainable long-term solution. It is mentioned that history has shown that regulations may not last long, and the incentives of different actors can change over time. These evaluation regimes can serve as a gradual transition toward a future with AGI, allowing more time for research in alignment and obtaining public and governmental consensus on the actions to be taken.

Marius Hobbhahn
Director of Apollo Research

Apollo wants to conduct AI model evaluations based on capabilities, using prompting and fine-tuning as main techniques. They believe that capability evaluations are more neglected than alignment evaluations.

In this context, they want to focus on deceptive alignment because (1) it is an instrumental capability for carrying out many actions that can cause large-scale harm, and (2) it is an overlooked issue. Operationalizing this capability involves, among other things, detecting if the model has situational awareness and the motivation to maintain its objectives.

Apollo proposes conducting evaluations during training for two reasons: (1) some capabilities may arise after a certain amount of training computation, and (2) there is a lack of understanding about how and why these capabilities emerge. Hobbhahn argues that evaluations during training are inexpensive and do not require pausing the training. He also believes that models are approaching the point of developing dangerous capabilities, and strict precautionary measures should be implemented, such as ensuring that a system can be shut down.

Evaluations during design also seem desirable, but there is less clarity on how to predict capabilities. Some ideas include (1) using computation-based scaling laws and (2) conducting qualitative tests to assess the model's improvement between two benchmarks. Hobbhahn suggests that training plans for models above a certain threshold should be approved by an authority.

Hobbhahn considers it important to develop standards based on general principles to guide the auditing processes. At the same time, these processes should feed into the standards created in an iterative process to refine both the standards and the audits.

The risks associated with conducting audits are discussed. Hobbhahn believes it is appropriate not to publish most of the results to avoid misuse of that information. However, he clarifies that for some evaluations, it makes sense to publish everything, while for others it doesn't. It is necessary to carefully consider the case before publishing.

It is also considered important to conduct evaluations in controlled environments to prevent leaks. Some ideas for achieving this include (1) using APIs, (2) performing the evaluation on the company's hardware, or (3) designing the evaluation for the company to execute in a process supervised by the evaluating organization.

Appendix 2 Literature Review

Summary of the literature review sections, in which sources were divided into three categories: Strategic research, policy research, and official documents.

1. Strategic Research

Strategic research in the field of artificial intelligence (AI) has gained significant importance in recent years due to the benefits that this technology can bring to society, as well as the risks and challenges associated with its development and use. In this regard, various specialized sources have addressed different aspects of this topic, providing valuable information and perspectives for the understanding and management of AI. The following are some of these sources:

Firstly, (European Commission. Directorate General for Communications Networks, Content and Technology & High-Level Expert Group on Artificial Intelligence, 2019) and (Scharre, 2019) highlight the importance of developing ethical guidelines to ensure trust in AI, in line with the values and fundamental rights of the European Union. On the other hand, the report by (Brundage, Avin, Clark, Toner, Eckersley, Garfinkel, Dafoe, Scharre, Zeitzoff, Filar, et al., 2018) (Brundage et al., 2020) emphasizes the need to anticipate and mitigate potential malicious uses of AI, as well as to foster international collaboration for this purpose.

Regarding the risks and benefits of AI, (Conn, 2015) highlights that while this technology can bring significant advances in areas such as health, transportation, or energy, it can also have negative effects on privacy, employment, or security. In this regard, (Yudkowsky, 2008) warns about the risk of AI as a global risk factor, while the report by (Hatzius et al., 2023) highlights the potential of AI to drive economic growth but raises concerns about the dangers it poses to jobs in various sectors globally, finding that up to 25% of jobs in the Eurozone are vulnerable. The detection and mitigation of emerging threats, such as automated influence operations targeting the general public, are discussed by (Goldstein et al., 2023).

On the other hand, (Ngo, 2020) and (Amodei et al., 2016) address the importance of aligning AI goals with human interests to avoid potential unintended consequences. They propose a series of measures such as safety helmets or sandboxes, control in design, rigorous experimentation and testing, transparency, and explainability. However, there are still major unknowns in the field that increase the unpredictability of AI systems. In line with this, the report by (AI Alignment 2018-19 Review - AI Alignment Forum, 2020) emphasizes the need to research and develop solutions to ensure AI safety and alignment.

In order to reduce the risks associated with AI, (Babcock et al., 2016) suggests a series of parameters such as secure design, which includes incorporating security mechanisms and minimizing vulnerabilities. Verification and validation involve testing and validating AGI before its implementation, aligning with (Ngo, 2020). Continuous monitoring and control are necessary to be able to control it in case it becomes dangerous or unexpected. Lastly, physical containment refers to the need to have physical measures in place to contain AGI if it becomes uncontrollable, such as "Kill Switches." The article (AI Alignment Forum, 2020) describes 11 proposals for building advanced and safe AI, highlighting the creation of a network of regulatory agencies, international cooperation in research, risk management, and policy formulation, anticipating risks from the design stages, incorporating "kill switch" mechanisms, transparency and explainability, verifiability through auditing and certification complying with international standards, establishing safety and ethical standards for AI developers, promoting education and public awareness and awareness of the risks and benefits of AI.

In terms of AI advancements, the article by (Mnih et al., 2015) highlights the potential of AI to achieve levels of control and learning comparable to humans through deep reinforcement learning. In this regard, the research by (Amodei & Hernandez, 2018) from OpenAI shows the exponential growth of computing power used in AI in recent years. Lastly, (Carlsmith, 2022) addresses the risk of AI as an agent of power and control, which could have existential consequences for humanity. In this regard, the report by (Scharre, 2019) (Brundage, Avin, Clark, Toner, Eckersley, Garfinkel, Dafoe, Scharre, Zeitzoff, Filar, et al.,

2018) emphasizes the importance of anticipating and regulating the use of AI-based "killer apps," which could have serious effects on national and global security.

In conclusion, strategic research in AI involves addressing different aspects, from ethics and trust to security and alignment with human interests. To achieve this, a multidisciplinary and collaborative perspective is necessary, taking into account both the benefits and risks of this technology, and working towards finding solutions to maximize its benefits and minimize its risks and challenges.

2. Policy Research

Artificial Intelligence (AI) is a rapidly developing technology that has generated interest in the political sphere. The implications of AI are vast, and therefore, several policy research efforts have been conducted in this field. The following are summaries of some of the most important research conducted in this area:

The article by (Brundage et al., 2020) focuses on the development of reliable AI. Trust in AI is essential for its acceptance in society. This work proposes a mechanism to support verifiable claims in AI development. AI verification can be done through certification, which is discussed in the article by (Cihon et al., 2021). Certification is a process that can help reduce information asymmetries in the ethical practice of AI. It proposes an AI certification framework to improve transparency and accountability in its development and use. The report emphasizes the need for independent and standardized evaluation of AI systems and the importance of disclosing information about the performance and safety of AI.

Security is another important concern in AI development. Building advanced and secure AI highlights the following points: creating a network of regulatory agencies, international cooperation in research, development, risk management, and policy formulation, anticipating risks from the design stage, incorporating "kill switch" mechanisms, transparency and explainability, verifiability through auditing and certification that complies with international standards, establishing security and ethical standards for AI developers, promoting education and public awareness of the risks and benefits of AI. These points align with the "Future Proof" report from the Center for Long-Term Resilience (Ord et al., 2021), the "AI Governance: A Research Agenda" report (Dafoe, 2018), and the "Policymaking in the Pause" report (Future Of Life Institute, 2023).

The proposal of "Auditing large language models: A three-layered approach" by (Mökander et al., 2023) specifically addresses the security of language models. The article describes a way in which language models can be audited in three layers: the input layer, the attention layer, and the output layer. This strategy can help identify and prevent the spread of misinformation or potentially dangerous information.

The role of cooperation in the responsible development of AI is addressed by (Askell et al., 2019). Cooperation is necessary to build responsible AI as the responsibility in AI development is shared by many stakeholders. The regulation of AI is another important topic discussed in the article by (European Commission, 2021). This article examines the AI standardization landscape, listing important standards such as ISO/IEC 23894:2020, IEEE P7003, ISO/IEC 30141:2019, NIST SP 800-53, IEEE 1291, and how they relate to the

European Commission's proposal for an AI regulatory framework. There is also a standard published this year, ISO/IEC DIS 42001.

The article by (Whittlestone & Clark, 2021) suggests that AI has great potential to impact society, and governments have an important role in ensuring it is developed responsibly and fairly. In this regard, it is proposed that governments monitor the development of AI. Recommendations are made regarding corporate governance of AI in the article by (Cihon et al., 2021), suggesting that companies adopt corporate governance practices to ensure AI is developed in the public interest.

Lastly, the article by (Tucker et al., 2020) addresses how data efficiency can positively affect society and AI governance in making better decisions. Additionally, improving data efficiency can help reduce costs and increase productivity. However, increasing data efficiency can have negative implications, such as increased surveillance, privacy violations, loss of freedom, and exacerbation of existing inequalities. It can lead to bias in decision-making algorithms, which can have negative consequences for marginalized or underrepresented social groups. Moreover, it can lead to a concentration of power in the hands of a few large corporations that control vast amounts of data. In such cases, governments need to take greater action to prevent these scenarios.

The report by (Siegmann & Anderljung, 2022) "The Brussels Effect and Artificial Intelligence: How EU regulation will impact the global AI market" discusses the impact of the EU's General Data Protection Regulation (GDPR) on global regulation of artificial intelligence (AI), highlighting the challenges posed by the extraterritorial application of GDPR and its interaction with other AI regulatory frameworks. It argues that the upcoming AI regulation in the EU is poised to have a global impact similar to that of the GDPR due to the EU's ability to influence global regulations through its market power and regulatory standards. The existence of a significant Brussels Effect may lead to stricter global regulation of AI, increasing the importance of helping shape the EU's AI regulatory regime. It is essential to ensure that EU AI regulation is future-ready and can adapt to a world of increasingly transformative AI capabilities.

3. Official documents

In recent years, several governments have presented documents and strategies related to artificial intelligence (AI). Here is a summary of the key points from six government documents addressing AI.

- "A Pro-Innovation Approach to AI Regulation" (Department for Science, Innovation & Technology, 2023) is a document from the British government addressing AI regulation. The document advocates for a pro-innovation approach that allows companies to experiment with AI without excessive regulation. The government acknowledges that AI has the potential to improve many aspects of life but also recognizes the need to establish certain regulations to ensure that AI is safe and reliable.

- The European Commission introduced the "Artificial Intelligence Act" in 2021, which sets harmonized rules for AI across the European Union. The legislation classifies AI into four categories: unacceptable AI, high-risk AI, limited-risk AI, and permitted use AI. High-risk AI will be subject to stricter safety and transparency requirements.

- The National Artificial Intelligence Strategy of Spain (SEDIA, 2020) (in effect), presented in 2020, emphasizes the need to invest in AI research and development to keep Spain at the forefront of the technology. The importance of public-private collaboration was also highlighted, along with the need to ensure that AI is used ethically and responsibly.

- The Spanish Strategy for Artificial Intelligence in R&D&I (Ministry of Science, Innovation, 2019) (complementary or non-effective) focuses on the application of AI in medical research and healthcare. The strategy sets specific goals, such as developing AI algorithms to improve early disease detection and enhance the efficiency of clinical trials.

- The Recovery, Transformation, and Resilience Plan of Spain (Government of Spain, 2021b) includes mentions in three specific areas, one of them being digital transformation, outlining significant investments in digital technologies, including AI, to be made until 2025. The plan states that investment in digital technology is crucial for economic recovery and the transformation of the Spanish economy.

- "España Digital 2026" (Government of Spain, 2021a) is a government strategy that sets a roadmap for the digital transformation of Spain over the next five years. The strategy highlights the importance of AI for the Spanish economy and establishes specific goals for its development, such as creating a national AI strategy and investing in research and development projects.

The analyzed government documents emphasize the importance of AI for the economy and social well-being. Governments recognize the need to establish regulations to ensure that AI is safe and reliable while fostering innovation and development in the field of AI. The significance of public-private collaboration and investment in research and development is also highlighted to stay at the forefront of technology.

Appendix 3 Actors Mapping

Understanding the development of the sandbox also involves understanding the structure of actors that is being developed in Europe on this issue and the allocation of functions that are taking place in the context of governance and regulation. To gather this information, a series of interviews were conducted with Spanish and international experts, as well as institutional actors related to the development of the sandbox.

The interviewees were selected based on the content of the report and their availability, given the global attention on AI development.

The content of the report is intended to address artificial intelligence from the historical opportunity of the publication of the EU Regulation on AI as the first law on this topic covering a total of 27 countries (while other countries like the USA have not yet developed one), and how this historical opportunity is unfolding in light of risks arising from the continuous use and expansion of AI knowledge that are exploited by a list of vectors taking advantage of system vulnerabilities, which become concrete threats. There are also a series of structural risks that transform cultural, social, and economic systems on a large scale worldwide. The report then analyzes the proposed European Regulation in relation to these vectors, threats, and risks, and how the sandbox constitutes a tool for testing proposals, regulations, and technology to enable the development of a safe market and industry for humanity.

Based on this structure, a series of experts and institutional actors were chosen to aid the understanding of each topic.

Regarding the experts, three Spanish experts, José H. Orallo, Toni Lorente, and Pablo Villalobos, as well as four experts of different nationalities, Charlotte Siegmann, Markus Anderljung, Samuel Hilton, and Risto Uuk, participated. The Spanish experts provided insights into the current state of AI in the global and local context, as well as the challenges that the European regulation faces in the development of the sandbox and the current state of the local industry in terms of technology and AI. On the global side, Charlotte Siegmann and Markus Anderljung were particularly helpful in understanding the European Regulation, its scope, and specifically the tests that it will have to face in order to promote governance and regulation of AI systems, and how this will have a global effect if done correctly. In terms of communications, Samuel Hilton helped to understand the challenges that different actors (including government, private sector, and academia) will face in effectively implementing the rules and recommendations that arise from the development of different policy, legal, and economic tools that make up AI. Lastly, but with great satisfaction, Risto Uuk helped us understand the global reach of AI and specifically how the sandbox will be a useful tool for testing institutions and capabilities, and how it is a great challenge from which valuable lessons can be derived, leading to an understanding of the dimension of the development of AI systems for humanity.

Regarding institutional actors, access to information has been more limited as the regulation is awaiting discussion by the three representative bodies of the European Union, and the sandbox has not yet formally started. There is still a degree of uncertainty regarding responsibilities and functions within the sandbox, which means that institutions that are currently working on it (such as the State Secretariat for Digitization and Artificial Intelligence, the Advisory Council on Artificial Intelligence, the National Cybersecurity Institute, among others) and those that have not yet been established (such as the Spanish Agency for Artificial Intelligence) are awaiting further definition of how the sandbox will function and what responsibilities each will assume within it.

Appendix 4 European Union Artificial Intelligence Act

The main starting point of the proposed law is the classification of AI systems according to the level of risk they entail. Specifically, the proposal is based on a hierarchy that distinguishes between unacceptable, high, limited, and minimal risks. The first two are the main focus of the regulation.

As part of the category of unacceptable risks, practices that pose a clear threat to security, livelihoods, and the rights of individuals are prohibited. Currently, three practices have been deemed unacceptable as they go against European values: altering human behavior to cause harm, assessing and classifying individuals based on their social behavior, and using real-time remote biometric identification systems in public spaces, except in cases of emergency.

On the other hand, high-risk systems are those with the potential to cause greater impact when deployed in critical sectors, including essential infrastructure, education, employment, essential public and private services, law enforcement, and border management. In this case, several requirements are imposed on the development and implementation of all products.

Firstly, providers of high-risk systems are required to establish, implement, document, and maintain a two-phase risk management system. Firstly, known and foreseeable risks must be identified and assessed both before and after commercialization. Risks can be considered "known" or "foreseeable" if the AI system developer should reasonably be aware of them. However, the Regulation does not currently clearly explain what constitutes "a reasonable level of diligence."

The second phase consists of reducing the detected risks to an acceptable level: providers must completely eliminate risks to the extent possible or, if not feasible, implement mitigation and control measures, as well as train users to make responsible use. Thus, the risk management system will be a process to be repeated until all identified risks become acceptable. The identification of unacceptable risks that cannot be reduced will result in the immediate cessation of the development and/or deployment of the AI system in question (Schuett, 2023b).

In parallel, providers will develop a quality management system to ensure that the development and verification of the AI system comply with the Regulation. Before market entry, developers must provide technical documentation that includes details of the system's design and architecture. Additionally, the training datasets must have followed governance guidelines regarding the choice of appropriate design, timely processing operations, and detection of potential deficiencies and biases.

Procedures aimed at enhancing cybersecurity and robustness, i.e., the system's resilience to alterations, will also be outlined. Transparency measures will be required, such as providing accessible instructions for use and, when applicable, informing the user that they are interacting with an AI system. Based on the documentation, predominantly internal evaluation procedures will be conducted. If the AI systems pass this examination, they will be endorsed by a conformity declaration drafted by the provider and made available to the authorities.

Throughout the period of use, systems must be supervised by humans who understand the capabilities and limitations of the model and can intervene in its operation if necessary.

Simultaneously, events (logs) occurring throughout the lifecycle will be automatically recorded to ensure traceability. In post-market monitoring, any severe incident or failure must be reported. In such cases, European market surveillance authorities will have the right to access data, documentation, and source code. When the operator is unable to take corrective measures, the authorities will also have the power to prohibit or restrict the marketing of the system.

For the implementation of the regulation, the EU advocates for the creation of controlled testing environments or sandboxes, which aim to identify and address potential issues in the application of the Regulation. These environments will be made available through calls for participation, allowing companies and organizations that wish to test new AI solutions to participate in them. The selected projects to integrate the sandboxes will be able to share information and knowledge, thus fostering collaboration and the exchange of experiences and best practices. Additionally, they will have access to guidance and expertise, providing a secure and controlled environment to test AI solutions before their market launch. The results of the conducted tests will contribute to the European Commission's efforts in effectively implementing the new Regulation and facilitating flexibility and adaptation of the rules to the real needs demanded by this technology.

In this context, the regulation mandates the assignment of national supervisory authorities and introduces the European Committee on Artificial Intelligence as the nexus for all state bodies. During the sandbox period, national authorities must submit annual reports to the Committee and the Commission, including results, lessons learned, and recommendations.

References

Agudo, U., & Matute, H. (2021). The influence of algorithms on political and dating decisions. *PLOS ONE*, *16*(4), e0249454. https://doi.org/10.1371/journal.pone.0249454

Aksela, M., Marchal, S., Patel, A., & Rosenstedt, L. (2022). *The security threat of AI-enabled cyberattacks* (p. 30). Finnish Transport and Communications Agency Traficom. https://www.traficom.fi/sites/default/files/media/publication/TRAFICOM_The_security_threat_of_AI-enabled_cyberattacks%202022-12-12_en_web.pdf

AMETIC. (2023). *MAPEO DEL ECOSISTEMA ESPAÑOL DE MICROELECTRÓNICA*. AMETIC. https://ametic.es/wp-content/uploads/2023/04/Mapeo_202304024_B.pdf

Anderson, H. S., Woodbridge, J., & Filar, B. (2016). DeepDGA: Adversarially-tuned domain generation and detection. *Proceedings of the 2016 ACM workshop on artificial intelligence and security*, 13-21.

Anderson, J. (2021, abril 15). *The dynamics of data accumulation*. Bruegel | The Brussels-Based Economic Think Tank. https://www.bruegel.org/blog-post/dynamics-data-accumulation

Artificial Intelligence: Guidelines for military and non-military use | News | European Parliament. (2020, octubre 12). https://www.europarl.europa.eu/news/en/press-room/20201209IPR93411/artificial-intelligence-guidelines-for-military-and-non-military-use

Baker, M. (2023). *Nuclear Arms Control Verification and Lessons for AI Treaties* (arXiv:2304.04123). arXiv. http://arxiv.org/abs/2304.04123

Bawden, D., & Robinson, L. (2020). Information Overload: An Introduction. En D. Bawden & L. Robinson, *Oxford Research Encyclopedia of Politics*. Oxford University Press. https://doi.org/10.1093/acrefore/9780190228637.013.1360

Bird, J., & Layzell, P. (2002). The evolved radio and its implications for modelling the evolution of novel sensors. *Proceedings of the 2002 Congress on Evolutionary Computation. CEC'02 (Cat. No.02TH8600)*, *2*, 1836-1841. https://doi.org/10.1109/CEC.2002.1004522

Blattner, L., & Nelson, S. (2021). How costly is noise? Data and disparities in consumer credit. *arXiv preprint arXiv:2105.07554*.

Bostrom, N., Dafoe, A., & Flynn, C. (2018). *Public Policy and Superintelligent AI: A Vector Field Approach*. https://nickbostrom.com/papers/aipolicy.pdf

Bou, C. P. (2023, marzo 9). *Los responsables del ciberataque al Hospital Clínic exigen un rescate a la Generalitat*. elperiodico. https://www.elperiodico.com/es/sanidad/20230309/ciberataque-ransom-house-hospital-clinic-rescate-ciberseguridad-84388678

Brundage, M., Avin, S., Clark, J., Toner, H., Eckersley, P., Garfinkel, B., Dafoe, A., Scharre, P., Zeitzoff, T., Filar, B., Anderson, H., Roff, H., Allen, G. C., Steinhardt, J., Flynn, C., hÉigeartaigh, S. Ó., Beard, S., Belfield, H., Farquhar, S., … Amodei, D. (2018). *The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation*.

https://doi.org/10.48550/ARXIV.1802.07228

Brundage, M., Avin, S., Wang, J., Belfield, H., Krueger, G., Hadfield, G., Khlaaf, H., Yang, J., Toner, H., Fong, R., Maharaj, T., Koh, P. W., Hooker, S., Leung, J., Trask, A., Bluemke, E., Lebensold, J., O'Keefe, C., Koren, M., … Anderljung, M. (2020). *Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims*. https://doi.org/10.48550/ARXIV.2004.07213

Buolamwini, J., & Gebru, T. (2018). Gender shades: Intersectional accuracy disparities in commercial gender classification. *Conference on fairness, accountability and transparency*, 77-91.

Carlsmith, J. (2022). *Is Power-Seeking AI an Existential Risk?* https://doi.org/10.48550/ARXIV.2206.13353

Castillo, C. del. (2022, junio 29). *España es el país con más ciberataques para robar contraseñas o datos bancarios*. elDiario.es. https://www.eldiario.es/tecnologia/espana-pais-ciberataques-robar-contrasenas-datos-bancarios_1_9129484.html

Clarke, S., Whittlestone, J., Maas, M., Belfield, H., Hernández-Orallo, J., & Heigeartaigh, S. Ó. (2021). *Submission of Feedback to the European Commission's Proposal for a Regulation laying down harmonised rules on artificial intelligence* [Feedback]. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Artificial-intelligence-ethical-and-legal-requirements/F2665626_en

Consejo. (2022). *Artificial Intelligence Act: Council calls for promoting safe AI that respects fundamental rights*. https://www.consilium.europa.eu/en/press/press-releases/2022/12/06/artificial-intelligence-act-council-calls-for-promoting-safe-ai-that-respects-fundamental-rights/

Dressel, J., & Farid, H. (2018). The accuracy, fairness, and limits of predicting recidivism. *Science advances*, *4*(1), eaao5580.

El Gobierno ya está invitando a grandes empresas al ensayo del Reglamento de la IA: Quiere resultados para noviembre. (2023, marzo 6). Business Insider España.

https://www.businessinsider.es/gobierno-ya-invita-empresas-ensayos-reglamento-ia-1209262

Eloundou, T., Manning, S., Mishkin, P., & Rock, D. (2023). *GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models* (arXiv:2303.10130). arXiv. http://arxiv.org/abs/2303.10130

EU AI Act needs clear safeguards for AI systems for military and national security purposes | ECNL. (2022, marzo 23). https://ecnl.org/news/eu-ai-act-needs-clear-safeguards-ai-systems-military-and-national-security-purposes

European Commission. Directorate General for Competition. (2019). *Competition policy for the digital era.* Publications Office. https://data.europa.eu/doi/10.2763/407537

Fruhlinger, J. (2022, agosto 31). *Stuxnet explained: The first known cyberweapon*. CSO Online. https://www.csoonline.com/article/3218104/stuxnet-explained-the-first-known-cyberweapon.html

Future Of Life Institute. (2021). *FLI position on the Proposal for a Regulation laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. https://ec.europa.eu/info/law/better-regulation/have-your-say/initiatives/12527-Inteligencia-artificial-requisitos-legales-y-eticos/F2665546_es

GAWER, A., MANNE, G., STUCKE, M., VARIAN, H., & BURNSIDE, A. J. (2016). *Big data: Bringing competition policy to the digital era*. https://www.oecd.org/competition/big-data-bringing-competition-policy-to-the-digital-era.htm

Goldstein, J. A., Sastry, G., Musser, M., DiResta, R., Gentzel, M., & Sedova, K. (2023). *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations* (arXiv:2301.04246). arXiv. http://arxiv.org/abs/2301.04246

Goodfellow, I., Bengio, Y., & Courville, A. (2016). *Deep learning*. The MIT Press.

Hatzius, J., Briggs, joseph, Kodnani, D., & Pierdomenico, G. (2023). The Potentially Large Effects of Artificial Intelligence on Economic Growth (Briggs/Kodnani). *Goldman Sachs Economic Research*. Goldman Sachs. https://www.key4biz.it/wp-content/uploads/2023/03/Global-Economics-Analyst_-The-Potentially-Large-Effects-of-Artificial-Intelligence-on-Economic-Growth-Briggs_Kodnani.pdf

Hwang, T. (2018). *Computational Power and the Social Impact of Artificial Intelligence* (arXiv:1803.08971). arXiv. http://arxiv.org/abs/1803.08971

INCIBE. (2023). *Red Team Aguas Misteriosas | INCIBE-CERT | INCIBE*. https://www.incibe.es/incibe-cert/blog/red-team-aguas-misteriosas

Ji, Z., Lee, N., Frieske, R., Yu, T., Su, D., Xu, Y., Ishii, E., Bang, Y., Dai, W., Madotto, A., & Fung, P. (2023). Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, *55*(12), 1-38. https://doi.org/10.1145/3571730

Königs, P. (2022). Artificial intelligence and responsibility gaps: What is the problem? *Ethics and Information Technology*, *24*(3), 36. https://doi.org/10.1007/s10676-022-09643-0

Krakovna, V., Uesato, J., Mikulik, V., Rahtz, M., Everitt, T., Kumar, R., Kenton, Z., Leike, J., & Legg, S. (2020). *Specification gaming: The flip side of AI ingenuity*. https://www.deepmind.com/blog/specification-gaming-the-flip-side-of-ai-ingenuity

McGregor, S., Paeth, K., & Lam, K. (2022). *Indexing AI Risks with Incidents, Issues, and Variants* (arXiv:2211.10384). arXiv. http://arxiv.org/abs/2211.10384

Miller, C. (2021, noviembre 11). *Throwback Attack: BlackEnergy attacks the Ukrainian power grid*. Industrial Cybersecurity Pulse. https://www.industrialcybersecuritypulse.com/threats-vulnerabilities/throwback-attack-blackenergy-attacks-the-ukrainian-power-grid/

Ministerio de Asuntos Económicos y Transformación Digital. (2022, junio 26). *El Gobierno de España presenta, en colaboración con la Comisión Europea, el primer piloto del sandbox de regulación de Inteligencia Artificial en la UE*. Mineco. https://portal.mineco.gob.es/es-es/comunicacion/Paginas/20220627-PR_AI_Sandbox

.aspx

Mökander, J., Axente, M., Casolari, F., & Floridi, L. (2022). Conformity Assessments and Post-market Monitoring: A Guide to the Role of Auditing in the Proposed European AI Regulation. *Minds and Machines*, *32*(2), 241-268. https://doi.org/10.1007/s11023-021-09577-4

Mökander, J., Schuett, J., Kirk, H. R., & Floridi, L. (2023). *Auditing large language models: A three-layered approach*. https://doi.org/10.48550/ARXIV.2302.08500

Nagesh, A. (2017, septiembre 18). Stanislav Petrov—The man who quietly saved the world—Has died aged 77. *Metro*. https://metro.co.uk/2017/09/18/stanislav-petrov-the-man-who-quietly-saved-the-world-has-died-aged-77-6937015/

Ngo, R., Chan, L., & Mindermann, S. (2022). *The alignment problem from a deep learning perspective*. https://doi.org/10.48550/ARXIV.2209.00626

Nguyen, T. T., Nguyen, Q. V. H., Nguyen, D. T., Nguyen, D. T., Huynh-The, T., Nahavandi, S., Nguyen, T. T., Pham, Q.-V., & Nguyen, C. M. (2022). Deep Learning for Deepfakes Creation and Detection: A Survey. *Computer Vision and Image Understanding*, *223*, 103525. https://doi.org/10.1016/j.cviu.2022.103525

O'Keefe, C., Cihon, P., Flynn, C., Garfinkel, B., Leung, J., & Dafoe, A. (2020). *The Windfall Clause: Distributing the Benefits of AI.* (p. 66). Centre for the Governance of AI Research Report. Future of Humanity Institute, University of Oxford. https://www.fhi.ox.ac.uk/windfallclause/

Omohundro, S. M. (2007). The nature of self-improving artificial intelligence. *Singularity Summit*, *2008*.

OpenAI. (2023). *GPT-4 Technical Report* (arXiv:2303.08774). arXiv. http://arxiv.org/abs/2303.08774

Ord, T., Mercer, A., & Dannreuther, S. (2021). *FUTURE PROOF: THE OPPORTUNITY TO TRANSFORM THE UK'S RESILIENCE TO EXTREME RISKS* (p. 52). The Centre for the Study of Existential Risk.

https://www.cser.ac.uk/media/uploads/files/Future_Proof_report_June_2021.pdf

Perez, F., & Ribeiro, I. (2022). *Ignore Previous Prompt: Attack Techniques For Language Models* (arXiv:2211.09527). arXiv. http://arxiv.org/abs/2211.09527

Rodríguez, B. (2022, septiembre 29). España Lanza el primer piloto de “AI regulatory sandbox” de la Unión Europea | Observatorio IA. *Observatorio IA de AMETIC*. https://observatorio-ametic.ai/regulacion-de-la-inteligencia-artificial/espana-lanza-el-primer-piloto-de-ai-regulatory-sandbox-de

Russell, S. J. (2019). *Human compatible: Artificial intelligence and the problem of control*. Viking.

Russell, S. J., Norvig, P., & Davis, E. (2010). *Artificial intelligence: A modern approach* (3rd ed). Prentice Hall.

Sadeghi, M., & Arvanitis, L. (2023). Rise of the Newsbots: AI-Generated News Websites Proliferating Online. *NewsGuard*. https://www.newsguardtech.com/special-reports/newsbots-ai-generated-news-websites-proliferating

Schippers, B. (2020). Artificial intelligence and democratic politics. *Political Insight*, *11*(1), 32-35.

Schuett, J. (2023). Defining the scope of AI regulations. *Law, Innovation and Technology*, *15*(1), 60-82. https://doi.org/10.1080/17579961.2023.2184135

Schwarzschild, A., Goldblum, M., Gupta, A., Dickerson, J. P., & Goldstein, T. (2021). *Just How Toxic is Data Poisoning? A Unified Benchmark for Backdoor and Data Poisoning Attacks* (arXiv:2006.12557). arXiv. http://arxiv.org/abs/2006.12557

Seger, elizabeth, Avin, S., Pearson, G., Briers, M., Heigeartaigh, S. Ó., & Bacon, H. (2020). *Tackling threats to informed decisionmaking in democratic societies Promoting epistemic security in a technologically-advanced world* (p. 112). The Alan Turing Institute, Centre for the Study of Existential Risk, University of Cambridge. https://www.turing.ac.uk/sites/default/files/2020-10/epistemic-security-report_final.pdf

SETELECO. (2022). *Plan Sectorial Vigilancia Mercado Teleco 2022-2026*. SETELECO, Gobierno de España. https://avancedigital.mineco.gob.es/equipos-telecomunicacion/Documents/Plan-Sectorial-Vigilancia-Mercado-Teleco-2022-2026-1.pdf

Sevilla, J., Heim, L., Ho, A., Besiroglu, T., Hobbhahn, M., & Villalobos, P. (2022). *Compute Trends Across Three Eras of Machine Learning* (arXiv:2202.05924). arXiv. http://arxiv.org/abs/2202.05924

Shah, R., Varma, V., Kumar, R., Phuong, M., Krakovna, V., Uesato, J., & Kenton, Z. (2022). *Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals* (arXiv:2210.01790). arXiv. http://arxiv.org/abs/2210.01790

Shavit, Y. (2023). *What does it take to catch a Chinchilla? Verifying Rules on Large-Scale Neural Network Training via Compute Monitoring* (arXiv:2303.11341). arXiv. http://arxiv.org/abs/2303.11341

Shevlane, T. (2022). *Structured access: An emerging paradigm for safe AI deployment*. https://doi.org/10.48550/ARXIV.2201.05159

Shoshitaishvili, Y., Bianchi, A., Borgolte, K., Cama, A., Corbetta, J., Disperati, F., Dutcher, A., Grosen, J., Grosen, P., Machiry, A., & others. (2018). Mechanical phish: Resilient autonomous hacking. *IEEE Security & Privacy*, *16*(2), 12-22.

Siegmann, C., & Anderljung, M. (2022). *The Brussels Effect and Artificial Intelligence* (p. 97). Centre for the Governance of AI.

Urbina, F., Lentzos, F., Invernizzi, C., & Ekins, S. (2022). Dual use of artificial-intelligence-powered drug discovery. *Nature Machine Intelligence*, *4*(3), 189-191. https://doi.org/10.1038/s42256-022-00465-9

Vincent, J. (2023, marzo 8). *Meta's powerful AI language model has leaked online—What happens now?* The Verge. https://www.theverge.com/2023/3/8/23629362/meta-ai-language-model-llama-leak-online-misuse

Whittlestone, J., & Clark, J. (2021). *Why and How Governments Should Monitor AI Development* (arXiv:2108.12427). arXiv. http://arxiv.org/abs/2108.12427

Zhang, Z., Ning, H., Shi, F., Farha, F., Xu, Y., Xu, J., Zhang, F., & Choo, K.-K. R. (2022). Artificial intelligence in cyber security: Research advances, challenges, and opportunities. *Artificial Intelligence Review*, 1-25.

Zwetsloot, R., & Dafoe, A. (2019, febrero 11). *Thinking About Risks From AI: Accidents, Misuse and Structure*. Lawfare. https://www.lawfareblog.com/thinking-about-risks-ai-accidents-misuse-and-structure