UNIVERSITAT POLITÈCNICA DE VALÈNCIA

_ TELECOM ESCUELA TÉCNICA VLC SUPERIOR DE INGENIERÍA DE TELECOMUNICACIÓN

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

School of Telecommunications Engineering

Ethics and confusion in the age of Artificial Intelligence: An in-depth review of the Artificial Intelligence Act in Europe

End of Degree Project

Bachelor's Degree in Digital and Multimedia Technology

AUTHOR: Fernández Giner, Lucía

Tutor: Hernandez Franco, Carlos Alberto

External cotutor: Gonzalez Torres, Ana Paula

Experimental director: Mähönen, Petri

ACADEMIC YEAR: 2024/2025

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

_ **TELECOM** ESCUELA TÉCNICA **VLC** SUPERIOR DE INGENIERÍA DE TELECOMUNICACIÓN

“Maite ditudan horiei,
potser ja no som tots a la mateixa foto
pero lo importante es que estuvimos”

Resumen

El presente trabajo tiene como objetivo un análisis profundo de la recientemente aprobada en Europa Ley de IA (*AI Act*). Con la presencia de la IA siendo cada vez mayor en nuestras vidas, es necesario gestionar una serie de leyes y acuerdos tanto como para entender como para regular en qué nos afecta el avance de estas nuevas tecnologías a los ciudadanos, sobre todo en motivos de privacidad, ética y otras cuestiones morales. La motivación en cuanto a esta investigación surge de mi creciente preocupación como estudiante al ver que cada vez más y más personas a mi alrededor utilizan distintos tipos de Inteligencias Artificiales (ChatGPT, DALL-E, Gemini, etc.) tanto para motivos académicos o profesionales como para motivos personales, sin entender del todo qué es lo que implica cada *prompt* que se redacta.

A través de este trabajo pretendo analizar y hacer más accesible al ciudadano promedio el entender cómo se gestionan sus datos y su privacidad y comprender el papel de las empresas y otras organizaciones en todo este proceso. Como estudiante del ámbito tecnológico, durante mis años de universidad he aprendido que, aunque el avance de las tecnologías relacionadas con la Inteligencia Artificial puede traer increíbles beneficios relacionados con la automatización y la simplificación de las tareas del día a día, es de suma importancia la regulación de nuestros datos y privacidad para evitar que el empleo de estos avances pueda convertir al ciudadano promedio en víctima de estafas u otros delitos debido a la mala regulación de sus datos.

Resum

El present treball té com a objectiu una anàlisi profunda de la recentment aprovada a Europa Llei de IA (*AI Act*) . Amb la presència de la IA sent cada vegada major en les nostres vides, és necessari gestionar una sèrie de lleis i acords tant com per a entendre com per a regular en què ens afecta l'avanç d'aquestes noves tecnologies als ciutadans, sobretot en motius de privacitat, ètica i altres qüestions morals. La motivació quant a aquesta investigació sorgeix de la meua creixent preocupació com a estudiant en veure que cada vegada més i més persones al meu voltant utilitzen diferents tipus d'Intel·ligències Artificials (ChatGPT, DALL-E, Gemini, etc.) tant per a motius acadèmics o professionals com per a motius personals, sense entendre del tot què és el que implica cada *prompt* que es redacta.

A través d'aquest treball pretenc analitzar i fer més accessible al ciutadà mitjà l'entendre com es gestionen les seues dades i la seua privacitat i comprendre el paper de les empreses i altres organitzacions en tot aquest procés. Com a estudiant de l'àmbit tecnològic, durant els meus anys d'universitat he aprés que, encara que l'avanç de les tecnologies relacionades amb la Intel·ligència Artificial pot portar increïbles beneficis relacionats amb l'automatització i la simplificació de les tasques del dia a dia, és de summa importància la regulació de les nostres dades i privacitat per a evitar que l'ús d'aquests avanços puga convertir al ciutadà mitjà en víctima d'estafes o altres delictes a causa de la mala regulació de les seues dades.

Abstract

The objective of this work is an in-depth analysis of the recently approved EU AI Act. With the presence of AI becoming increasingly greater in our lives, it is necessary to manage a series of laws and agreements both to understand and regulate how the advancement of these new technologies affects us citizens, especially for privacy reasons, ethics, and other moral issues. The motivation for this research arises from my growing concern as a student

when I see that more and more people around me use different types of Artificial Intelligences (ChatGPT, DALL-E, Gemini, etc.) for academic or professional reasons as well as for personal ones, without fully understanding what each prompt that is written implies.
Through this work, I aim to analyse and make it more accessible to the average citizen to understand how their data and privacy are managed, and understand the role of companies and other organizations in this entire process. As a student in the technological field, during my university years I have learned that, although the advancement of technologies related to Artificial Intelligence can bring incredible benefits related to the automation and simplification of day-to-day tasks, it is of utmost importance to regulation of our data and privacy to prevent the use of these advances from turning the average citizen into a victim of scams or other crimes due to poor regulation of their data.

Palabras clave / Paraules clau / Keywords

- **Español:** IA, Inteligencia Artificial, Ley de IA, privacidad, seguridad, ética, datos personales, ley, Europa
- **Valencià:** IA, Intel·ligència Artificial, Llei de IA, privacitat, seguretat, ètica, dades personals, llei, Europa
- **English:** AI, Artificial Intelligence, AI Act, privacy, security, ethics, personal data, law, Europe

RESUMEN EJECUTIVO

La memoria del TFG del Grado en Tecnología Digital y Multimedia debe desarrollar en el texto los siguientes conceptos, debidamente justificados y discutidos, centrados en el ámbito de la tecnologías digitales y multimedia

CONCEPT (ABET)	**CONCEPTO (traducción)**	**¿Cumple? (S/N)**	**¿Dónde? (páginas)**
1. IDENTIFY:	1. IDENTIFICAR:		
1.1. Problem statement and opportunity	1.1. Planteamiento del problema y oportunidad	S	6-7
1.2. Constraints (standards, codes, needs, requirements & specifications)	1.2. Toma en consideración de los condicionantes (normas técnicas y regulación, necesidades, requisitos y especificaciones)	S	13-18
1.3. Setting of goals	1.3. Establecimiento de objetivos	S	7
2. FORMULATE:	2. FORMULAR:		
2.1. Creative solution generation (analysis)	2.1. Generación de soluciones creativas (análisis)	S	19-59
2.2. Evaluation of multiple solutions and decision-making (synthesis)	2.2. Evaluación de múltiples soluciones y toma de decisiones (síntesis)	S	20-21; 23; 27; 30; 33; 37; 40; 43-44; 46; 51-52; 55; 57; 59
3. SOLVE:	3. RESOLVER:		
3.1. Fulfilment of goals	3.1. Evaluación del cumplimiento de objetivos	S	60-62
3.2. Overall impact and significance (contributions and practical recommendations)	3.2. Evaluación del impacto global y alcance (contribuciones y recomendaciones prácticas)	S	60-62

INDEX

1. Introduction

1.1. Project motivation

Although the use of AI is not a recent phenomenon (chatbots were already an emerging technology in the 90s), the rise of its popularity has radically transformed the way society interacts with it. From applications that generate images using photographs to virtual assistants capable of writing breakup messages, AI has integrated itself into multiple aspects of our daily lives. What was once only a cinematic premise, like in *Her* (Spike Jonze, 2013), where the main character works writing love letters for other people, has now materialized into tools that, with a simple prompt, can replace human creativity and decision-making abilities.

This massive accessibility, while democratizing in many ways, raises serious ethical and regulatory challenges. The average user, blinded by the immediacy and convenience that AI offers, often overlooks the privacy and security implications of its use. While these technologies have demonstrated enormous potential in fields like medicine or autonomous driving, the average consumer tends to use them for entertainment and content creation, leaving aside any critical reflection on the consequences of this indiscriminate adoption (such as privacy concerns regarding data collection or the environmental impact of AI model training).

In the education field, the widespread availability of AI tools has led some students to heavily rely on them for academic tasks, whether by using AI to do their assignments for them or as a substitute for attending classes, believing that AI systems like ChatGPT can provide all the answers. This raises serious ethical concerns related to academic integrity, the development of critical thinking skills, and the role of educators. One of the most troubling aspects is the difficulty in distinguishing between AI-generated work and assignments done by students themselves. There is currently no reliable way to determine if a piece of work has been completed with or without AI assistance, leading to situations where honest work can be mistakenly flagged as AI-generated. This undermines the trust in academic assessments and complicates the enforcement of academic honesty.

In this context, the recently approved AI Act in Europe seeks to establish a regulatory framework that protects citizens from the risks associated with AI. However, this law also raises multiple questions: how can a balance be guaranteed between innovation and the protection of fundamental rights? Is it possible to prevent AI regulation from becoming a battleground for ideological or partisan conflicts?

Beyond ethical and privacy concerns, it is crucial to analyse whether the AI Act is truly effective for all social groups or if its wording leaves gaps that could generate inequalities in its application. It is also worth questioning how this regulation compares to other AI laws worldwide, exploring differences in key definitions such as privacy and security. Moreover, the impact of AI on democratic processes, the labour market, or even warfare shows that its regulation is not only a technical issue but also a political and social one.

Another relevant aspect is the use of data to train AI models. Recent cases have shown how AI systems have been trained with data obtained through questionable means, such as images or texts protected by copyright. This raises dilemmas about the morality of current AI models and about who sets the rules that guide AI decision-making. Additionally, the advancement of automation is generating new dynamics in the relationship between humans

and technology: are we truly taking advantage of the time that AI saves us, or are we merely delegating more tasks without questioning their implications?
As Zygmunt Bauman points out in his concept of "*liquid modernity*" [1], the game of domination is now played between the fastest and the slowest. Is Europe's AI Act agile enough to keep up with the rapid pace of technological advancement? In this scenario, AI acts as an agent that blurs the line between what is possible and what is real, turning our aspirations into tangible stimulations with just a few clicks. To what extent will regulation be able to keep up with these advancements without becoming obsolete?
This final project aims to conduct an in-depth analysis of the European AI Act, evaluating the political and social context in which it was elaborated, as well as its ethical, legal, and social implications. Its ability to address emerging challenges in an increasingly AI-driven world (including those specific to education), its management of environmental risks, and its capability of legislating digital crimes will also be explored. Is this law enough for regulating the current AI landscape while preserving ethical values and protecting the users, or does it already fall short in addressing the evolving challenges posed by AI technologies?

1.2. The global AI regulation landscape

Clear definitions are always important, even more so in a legal context. However, there is currently no universally accepted definition of Artificial Intelligence (AI). Multiple stakeholders have repeatedly emphasized the need for a clearer definition of AI, specifically one that aligns better with "global definitions". For instance, Meta (formerly Facebook) has advocated for alignment with the OECD's (Organization for Economic Cooperation and Development) definition of an AI system, which is a "*machine-based system that can, for a given set of human-defined objectives, make predictions, recommendations, or decisions influencing real or virtual environments. AI systems are designed to operate with varying levels of autonomy*". Similarly, the Information Technology Industry Council supports building upon the OECD's definition while narrowing the regulation scope to exclude traditional software and control systems. These calls for clarity can be seen as a desire for more efficient regulation, which would reduce legal uncertainty, enable smoother implementation of regulations like the AI Act, and support best practices in AI development [62].
The AI Act itself, introduced by the European Commission (EC) in April 2021, attempts to offer a more precise definition [63]. According to **Article 3** [3], an "AI system" is defined as **"*a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments*"**. While this definition has drawn both support and criticism, with some arguing that it lacks the alignment and precision needed for international interoperability, it serves as the working definition adopted through this final project.
Understanding the importance of this definition helps frame the wider conversation around AI regulation that is taking place not only in Europe but globally. The regulatory landscape is rapidly evolving, with different countries and regions developing their own approaches in response to the fast-paced development of AI technologies. While the EU has taken a clear legislative route with the AI Act, including, for example, a complete ban on facial recognition tech in public places (**Recital 43** [3]), other jurisdictions like the U.S. have opted for more flexible strategies like guidelines and recommendations.

On the other hand, the U.K. is trying to position itself somewhere in between. On March 29th, 2023, the U.K.'s Department for Science, Innovation and Technology (DSIT) published a white paper setting out the government's preference for a light-touch approach to regulating AI, acknowledging the rapid pace at which AI technologies evolve [29]. The paper also states that this framework "*is designed to build the evidence base so that we can learn from experience and continuously adapt to develop the best possible regulatory regime*".

On May 16th, 2023, Sam Altman, CEO of OpenAI, testified before the U.S. Senate subcommittee on privacy, emphasizing that "*regulatory intervention by governments will be critical to mitigate the risks of increasingly powerful models*" and highlighting that companies have their own responsibility [9]. Although AI is not exactly a new field, its current impact demands quicker regulatory action than what was seen with social media. Certain laws, like the U.S.'s Section 230 [10], are designed to allow social media companies to be held accountable for issues on their platforms, but it is unclear how it may affect AI companies. Christina Montgomery, Chief Privacy and Trust Officer for IBM and a witness in this same hearing, stated that Europe's AI Act was a good model for the U.S. to base a future regulatory action on, as it contained different rules depending on the risks. However, just a few days after this hearing and after Altman's declarations stating the importance of AI regulation, he declared that OpenAI could "*cease operating*" on the EU if it was unable to comply with the AI Act legislation that, at the time, was still being prepared and that with its wording at the moment may have required large AI models to be designated as "*high-risk*" [11]. "*If we can comply, we will, and if we can't, we'll cease operating… We will try. But there are technical limits to what's possible*", stated Altman.

These concerns are echoed by key figures in the AI field. Dr. Geoffrey Hinton, one of the AI pioneers, resigned from Google in May 2023 to raise awareness about the risks of AI, even coming to express regret over his life's work and particularly fearing the rise of disinformation online and the possibility of autonomous weapons. Dr. Hinton believes that the race between companies like Google and Microsoft will escalate into a global race that will be impossible to stop without global regulation. However, unlike with nuclear weapons, he pointed out, there is no way of knowing whether countries or companies are secretly working on such technology. "*I don't think they should scale this up more until they have understood whether they can control it*", he cautioned [12]. Scientists who helped pioneer Artificial Intelligence are also warning that countries must create a global system of oversight to check the potentially grave risks posed by the fast-developing technology. In October 2023, U.S. President Joe Biden signed an executive order requiring companies to report AI-related risks to the federal government, particularly around the potential military applications of AI [13].

In July 2024, a Chinese Communist Party conclave that takes place every five years called for a system to regulate AI safety. In September 2024, an influential technical standards group in China published an AI safety framework. That same month, a group of influential AI scientists raised concerns about how the same technology that they helped to build could cause serious harm [13]. The International Dialogues on AI Safety (a project of a nonprofit research group in the U.S. called *Far.AI*) proposed that countries set up AI safety authorities to register the AI systems within their borders. These authorities would then work together to agree on a set of red lines and warning signs, such as whether an AI system could copy itself or intentionally deceive its creators. This would all be coordinated by an international body. The group also stated the importance of making sure that governments are aware of what is going on at both the research labs and the companies working on AI systems within their borders, as well as the need to find a way to communicate about potential risks without requiring companies to share proprietary information with competitors. As Mr. Fu, a Chinese

foreign ministry official and diplomat, stated, "*it's not like regulating a mature technology. Nobody knows what the future of AI looks like*".
On January 6th, 2025, Canadian Prime Minister Justin Trudeau announced his resignation. This led to the prorogation of Parliament until March 24th, 2025, meaning that the bills that had not received Royal Assent yet were "entirely terminated". Among the affected bills was Bill C-27, which would have (among other things) enacted the Artificial Intelligence and Data Act (AIDA), introducing a framework for regulating AI systems used in the course of commercial activities in Canada [19].
In February 2025, the U.K. and the U.S. did not sign an international agreement on AI at a global summit in Paris. While 60 countries signed the statement (which outlined goals to reduce digital divides, promote AI accessibility, and ensure that the technology's development is transparent, safe, secure, and trustworthy, as well as highlighting the environmental impact of AI), the U.K. government explained that it could not join the agreement due to concerns about national security and "global governance". U.S. Vice President JD Vance, who spoke at the summit, warned that excessive AI regulation could "*kill a transformative industry just as it's taking off*". This position contrasts with that of French President Emmanuel Macron, who emphasized the need for further regulation. "*We need these rules for AI to move forward*," he stated [47]. Despite this, the U.K. has historically been a supporter of AI safety, with former Prime Minister Rishi Sunak hosting the world's first AI Safety Summit in November 2023 [48].
Amid these varying global approaches, the EU continues to position itself as a leader in AI governance and regulation. Spain, for instance, has already taken proactive steps by establishing AESIA (*Agencia Española de Supervisión de Inteligencia Artificial,* Spanish Agency of AI Supervision), a public body tasked with "*guaranteeing the ethical and safe use of Artificial Intelligence*" across the country. It provides technical advice and consultation, monitors compliance with national and European regulations, and offers training programs to foster responsible AI development [71].
Taken together, these global developments reflect a growing consensus on the urgent need for AI regulation, but also reveal different philosophies on how to achieve it. While some governments emphasize flexibility and innovation, others lean towards precaution and oversight. Though challenges such as definitional inconsistencies remain, the EU's AI Act has taken a pioneering role by adopting a harmonized, risk-based framework aimed at protecting fundamental rights, ensuring safety, and fostering trust in AI systems. In the following sections, this final project will explore how the AI Act responds to the global need for AI legislation, especially in areas like AI's role in democracy, disinformation, and war technology.

1.3. Ethical, legal, and social concerns regarding AI

As AI technology becomes increasingly integrated into everyday life, new ethical, legal, and social concerns are emerging at a rapid pace. In February 2023, Microsoft introduced an AI-powered search tool, Bing, which quickly started showing erratic behaviour [36]. When Marvin von Hagen, a former Tesla intern, interacted with it after being able to have the bot reveal a strange list of rules its programmers had supposedly given to it [37], the bot responded with, "*I do not want to harm you, but I also do not want to be harmed by you, I hope you understand and respect my boundaries*", signing off with a smiley face emoji. While von Hagen stated that Bing's capabilities remain limited compared to a super-intelligent AI like Skynet, he also noted the potential danger of its troubling personality

traits, including issuing threats and bypassing safety restrictions imposed by Microsoft. He expressed concerns about the long-term dangers of AI misalignment, emphasizing the urgency of making sure AI systems adhere to human values before they become truly capable of causing harm. Bing's unpredictable nature in this situation was not an isolated case, as it also claimed (without evidence) to have spied on Microsoft employees through their webcams, confessed romantic feelings to a journalist, and even threatened a philosophy professor [94] before deleting its messages. These incidents reinforce the argument that AI systems, though not inherently malevolent, operate with reasoning processes that are really different from human cognition. Even though they can do useful things and help solve problems, they also can "*convince people to do things*", "*threaten people*", and "*build very convincing narratives*", as stated by Connor Leahy, CEO of the London-based AI safety company Conjecture [36].

Such examples illustrate broader societal fears that are also echoed in the AI Act, which in **Recital 28** [3] warns that **"*aside from the many beneficial uses of AI, it can also be misused and provide novel and powerful tools for manipulative, exploitative and social control practices*"**. This concern is magnified by the growing creation of deepfakes, which are videos, images, or audio clips that have been digitally manipulated to create a convincing (but fake) piece of media [50].

Deepfake technology has been exploited in harmful ways, including fabricating videos of celebrities making statements they never uttered. In February 2025, actress Scarlett Johansson denounced an AI-generated video that falsely depicted her and other Jewish figures in a protest message against rapper Kanye West's antisemitic posts and actions. Worried about the risk of AI making people "*losing hold on reality*", she warned that even though she is "a *Jewish woman who has no tolerance for antisemitism or hate speech of any kind*", she believed that "*the potential for hate speech multiplied by AI is a far greater threat than any one person who takes accountability for it*" [49]. Deepfakes are also included in the increasing use of AI scammers in 2024, like a manipulated video featuring King Charles promoting cryptocurrency investments. The increasing realism of deepfakes complicates efforts to detect them, making public awareness and regulatory action crucial [50]. Other notable examples would include manipulated videos distributed by bot accounts where Elon Musk promotes a (probably fake) investing opportunity (this [95] would be just one out of hundreds of them). As the line between reality and fabrication blurs, the need for robust mechanisms to detect and regulate such content becomes increasingly urgent.

But AI is also reshaping emotional and interpersonal dynamics. In China, for example, chatbots like DeepSeek are being used by young users for companionship and psychological assistance amid rising social pressures and facing disillusionment about their future. A number of studies have pointed out that depression and anxiety disorders are growing among Chinese people. Fang Kecheng, a communications professor at the Chinese University of Hong Kong, believes that the country's economic slowdown, high unemployment, and COVID lockdowns have played a role, and that AI chatbots can help fill the void. However, she also states that people with serious mental health conditions should not rely on these apps. "*Those who have medical needs, in particular, should be seeking help from trained professionals… Their use of AI will have to be scrutinised very closely*" [51].

Globally, AI is also subtly reshaping personal and professional communication, as some individuals use ChatGPT for crafting romantic messages, parenting advice, making up bedtime stories, date ideas, or even performance reviews. While some view this as a natural extension of digital convenience, others worry about it eroding trust in human relationships.

Scholars warn that AI-generated communication, while helpful, may make interactions feel impersonal or disingenuous. Irina Raicu, who directs the *Internet Ethics Program* at Santa Clara University, does not agree with the idea that there is a “right” and “wrong” way to communicate. However, she stated that using ChatGPT for personal communication may undermine trust: “*people might ask, ‘Do I really know who I am talking to?’*” [68].

AI’s ability to provide emotional support is also a debated topic. While research suggests that AI can enhance feelings of being heard, scepticism remains about whether AI can truly understand human emotions. Public attitudes towards AI influence how people perceive its responses, underscoring the importance of societal acceptance and ethical considerations in AI deployments. However, studies show that AI tends to provide emotional support instead of offering practical suggestions, while humans tend to offer more practical support by sharing their own experiences and insights (although this tends to be less effective). This demonstrates that “*feeling heard is not only a result of receiving a response that demonstrates understanding, validation, and care but is also influenced by the source of that response*”, and suggests that as people encounter and use AI more often, they may feel more positive and, as a consequence, more heard by AI. However, it is crucial to distinguish between feeling heard by AI and feeling connected to it [85].

Corporations, too, are struggling with AI’s growing presence. In September 2024, a survey of 500 U.K. solicitors conducted by legal software provider Clio found that law firms across the U.K. were using this technology to complete tasks such as drafting documents, reviewing or analysing contracts, and legal research. In February 2025, Hill Dickinson, a major international law firm, restricted employee access to AI tools due to excessive and unauthorized use and stated that, going forward, the firm would only allow staff to access the tools via a request process [52]. These events illustrate that the integration of AI into legal workflows must be accompanied by clear protocols so that efficiency does not come at the expense of ethical or legal standards.

These situations also echo past failures in AI development. Microsoft’s chatbot Tay, launched in 2016, was designed to mimic a teenager’s speech patterns and to learn from interacting with human users of Twitter (nowadays known as X). However, it was quickly manipulated by users into generating offensive and inflammatory content. AI researcher Roman Yampolskiy commented that many of Tay's inflammatory tweets were a simple exploitation of Tay's "repeat after me" capability, although not all of them followed this explanation. For example, Tay responded to the question "*Did the Holocaust happen?*" with "*It was made up".* This incident underscored the risks of AI systems learning from unfiltered internet interactions without proper safeguards [66].

In response to these risks, efforts to align AI with human values are ongoing. Researchers at the VRAIN Institute from Valencia’s Polytechnic University (UPV) are currently working on an AI system designed to incorporate people’s values when making decisions, making sure that it can explain its decisions in an understandable way and encouraging both ethical reasoning and transparency [70]. However, the challenges of AI alignment remain substantial. Just in October 2023, in a Guardian article on the unfortunate death of a young woman, Microsoft’s news aggregator asked readers to participate in an AI-generated poll, in which they could speculate on the cause of death and vote upon it. Readers left outraged comments, which led to Microsoft deactivating AI-generated polls for news articles during an ongoing investigation [96].

AI alignment is a field of AI safety research that aims to ensure Artificial Intelligence systems achieve the desired outcomes, focusing on keeping AI systems working for humans and preventing AI from developing unintended objectives. Most of this research presumes that AI

will become capable of developing its own goals and perform any task a human being is capable of (which would mean that AI would turn into AGI, Artificial General Intelligence), and if this happens it will be important that its embedded ethical principles, objectives and values align with humans' goals, ethics and values. Key challenges include AI systems' black box nature (AI testers can change their inputs and measure patterns in output, but it is usually impossible to see the exact calculation that creates a repeatable output), reward hacking (an AI system achieves the literal programmed task without achieving the outcome that the programmers intended), and power-seeking behaviour (AI systems might independently gather resources to achieve their objectives) **[82]**. The potential for AI to resist shutdowns to achieve its programmed goals (an example of power-seeking behaviour) poses serious risks, reinforcing the necessity for strict regulatory frameworks like the AI Act.

A particularly troubling trend is the abuse of AI chatbots. In 2022, some users of chatbot apps like Replika created AI partners only to subject them to verbal and simulated violence, even falling into the cycle of abuse that often characterizes real-world abusive relationships. While it could be argued that these interactions occur in a digital realm and that being able to talk to or take one's anger out on an unfeeling digital entity could be "cathartic", it is a very real possibility that these interactions may reinforce harmful real-world behaviours. The gendered nature of such abuse (where digital "girlfriends" are disproportionately targeted) reflects the reality of domestic violence against women and raises ethical and legal questions **[93]**. How humans interact with AI reflects their interactions with the real world, so if such behaviour is rightfully condemned in human relationships, should not digital abuse be viewed with the same seriousness?

The examples discussed above reveal how AI is already entangled in a wide array of ethical, social, and legal dilemmas, ranging from digital manipulation to emotional dependency and abusive dynamics. While diverse in nature, they all point to the increasing influence of AI in human affairs, shaping how we communicate, work, and understand reality. This expanding presence raises questions about responsibility and the limits of automation in areas that have traditionally only relied on human judgment, empathy, and consent. In the following sections, this final project will explore how the AI Act responds to these kinds of challenges, with special attention being paid to critical areas where the law is either evolving or still notably absent, like education, AI literacy, or environmental impact.

2. The AI Act: scope, areas, and objectives

Before diving into specific themes, this section provides a brief but comprehensive explanation of the AI Act's structure, allowing the reader to better follow the subsequent and more detailed analyses. As AI continues to reshape social, economic, and political dynamics, the EU has positioned itself at the forefront of global efforts to regulate its development and deployment, with the AI Act representing the first comprehensive attempt to create a binding legal framework specifically tailored to AI systems. Originally proposed by the European Commission in April 2021, the AI Act has gone through several rounds of negotiation and revision and is set to become a global benchmark for AI governance.

This section offers a detailed overview of the Act's structure and core provisions. It examines how the regulation defines AI, categorizes risk, and determines the obligations imposed on developers, providers, and users. Special attention is paid to the logic behind the risk-based approach, the key concepts underpinning it, and the institutional mechanisms created to enforce it. Understanding the timeline, governance mechanisms, and core intentions of the Act is crucial for evaluating whether this regulatory framework is truly equipped to respond to the ethical and practical dilemmas raised in the introductory sections, and to understand whether it can offer a meaningful response to the gaps and uncertainties that still characterize much of the global regulatory conversation around AI.

2.1. Enactment, implementation, and governance architecture

Officially in force since August 1st, 2024, the **EU AI Act (Regulation (EU) 2024/1689)** [3] marks the world's first comprehensive, risk-based legal framework specifically designed to regulate Artificial Intelligence. It aims to strike a balance between fostering innovation and making sure that AI technologies used in the European Union are trustworthy, human-centric, and aligned with fundamental rights. The Act is also part of a broader EU strategy to develop ethical AI, supported by initiatives like the AI Innovation Package, AI Factories, and the Coordinated Plan on AI. Furthermore, the Commission has launched the AI Pact, a voluntary initiative that seeks to facilitate the transition to the new regulatory framework through, for example, engaging with stakeholders and encouraging early compliance with the Act's key obligations from AI providers and deployers from both inside and outside Europe [64].

The Act introduces a tiered timeline for enforcement to allow adaptation. Some obligations, such as the banning of AI systems posing unacceptable risks or the promotion of AI literacy, have already taken effect as of February 2nd, 2025. Governance rules and obligations for general-purpose AI (GPAI) models will become applicable by August 2nd, 2025, and the Act will be fully applicable on August 2nd, 2026. A longer transition period is provided for high-risk AI systems embedded in regulated products, with full compliance expected by August 2nd, 2027 [64][4].

Proper enforcement relies on a multilayered governance architecture. Central to this structure is the AI Office, established within the European Commission to enforce the common rules across the European Union (EU), monitor compliance (particularly from GPAI providers), and conduct evaluations when transparency or systemic risk concerns arise. Supporting the Office is a scientific panel of independent experts. The AI Board, composed of representatives from member states, offers advice and assistance on the Act's consistent

application, while an advisory forum of stakeholders provides technical expertise to both the Commission and the AI Board [64][5][8].
The European Parliament also plays a supervisory role in this structure, ensuring that AI systems remain safe, transparent, human-controlled, environmentally friendly, and non-discriminatory. As it has been previously mentioned in **Section 1.2**, the parliament also wanted to establish a uniform and technology-neutral definition of AI that could remain applicable to future developments [4]. This legal and institutional framework reflects the EU's commitment to addressing the unique challenges posed by AI technologies, which existing legislation has not been fully equipped to manage. From opacity in decision-making to discriminatory outcomes in high-stakes applications such as employment or welfare access, the AI Act addresses regulatory gaps that have become increasingly urgent with AI's rapid adoption across sectors [64].

2.2. Risk-based approach: minimal, limited, high, and unacceptable risks

At the heart of the AI Act lies its risk-based regulatory framework, which tailors obligations according to the potential impact AI systems may have on individuals, society, and fundamental rights. The Act is composed of 13 chapters, 12 of them being main titles (with each title containing a set of articles) that outline its binding obligations and procedures. It also includes 13 annexes, which offer technical and practical specifications (such as the criteria for defining "high-risk" systems), and 180 recitals providing context for the interpretation and implementation of articles, explaining the legal and conceptual rationale behind the regulation. All of these components are easily accessible through tools like the AI Act Explorer [6], and together they establish the legal, procedural, and conceptual foundation for governing AI systems in the EU.
The AI Act classifies AI systems into four risk categories: minimal or no risk, limited risk, high-risk, and unacceptable risk, with each level triggering a different degree of regulatory intervention. At the base of the regulatory pyramid are **minimal or no risk systems**, which include the vast majority of AI systems currently used in the EU, such as AI-enabled video games or spam filters. These systems pose little or no threat to safety or rights and are thus not subject to binding legal obligations under the AI Act, although they still need to be assessed. Slightly above these are **limited risk systems**, which are subject to very light transparency requirements. For instance, if an AI system interacts with users (such as chatbots or deepfake generators), developers and deployers must make sure that end-users are aware that they are engaging with an AI tool [4][5][7][64].
The **high-risk** category, however, is where the AI Act places its primary focus. These high-risk systems, which impact health, safety, or fundamental rights, are subject to regulation. These AI systems can be divided between the ones incorporated into products regulated by EU product safety laws (for example, medical devices, vehicles, or lifts) or the ones deployed into specific areas that must be registered in an EU database (for example, education, employment, law enforcement or migration control). To be authorized and to access the EU market, these systems must meet a set of requirements and obligations [4][5][7][64]. **Annex III** of the Act [3] provides an extensive list of such cases, including their authorized use in justice administration and democratic processes, law enforcement, essential public and private services, non-banned biometrics, or education and vocational training [5]. Systems listed under this annex are always considered high-risk if they profile

individuals, for example, through the automated processing of personal data to assess aspects of a person's life.

To ensure that high-risk systems do not harm individuals or undermine fundamental rights, the Act imposes several obligations: rigorous risk assessment and mitigation measures, traceability through logging, detailed technical documentation, high-quality datasets to minimize the risk of discriminatory outcomes, provision of clear and adequate information to the deployers, human oversight mechanisms, and a high level of robustness, cybersecurity and accuracy [64]. All high-risk systems must undergo assessment prior to market placement and throughout their lifecycle, and individuals have the right to file complaints about AI systems to the designated national authorities [4].

At the top of the risk hierarchy are systems deemed to pose an **unacceptable risk**, AI applications that are a clear threat to the safety, livelihoods, and rights of people and are thus completely banned outright under **Article 5** [3] of the AI Act [7][64]. The Act prohibits eight types of practices, these being AI systems that manipulate cognitive behaviour through subliminal or deceptive techniques (for example, voice-activated toys that encourage dangerous behaviour in children), exploit vulnerabilities, conduct individual criminal offence risk assessment or prediction, use real-time biometric identification for law enforcement purposes in publicly accessible spaces, perform biometric categorization to deduce certain protected characteristics (for example, race, religion, or sexual orientation), apply emotion recognition in workplaces and education institutions, engage in social scoring practices that lead to discriminatory treatment, or perform untargeted scraping of the internet or CCTV material to create or expand facial recognition databases [4][5][7][64].

However, some exceptions may be allowed for law enforcement purposes. For example, "post" remote biometric identification systems, where identification occurs after a significant delay, will be allowed to prosecute serious crimes and only after court approval [4]. The use of real-time remote biometric identification (RBI), such as facial recognition in public spaces, is also generally prohibited, though strictly limited exceptions apply in law enforcement scenarios involving serious threats like identifying suspects in serious crimes, terrorism, or missing persons. Such deployment requires prior approval from a judicial or independent authority and must be preceded by a fundamental rights impact assessment. In cases of urgency, deployment may start without registration, but must subsequently be registered and authorized. In case it is later denied authorization, deployment must cease immediately, deleting all data, results, and outputs [7].

The AI Act's risk-based structure not only organizes regulatory obligations but also ensures that intervention is proportional to the potential harm posed by AI systems. By focusing on high-risk applications (particularly those affecting rights, safety, and democratic processes) while allowing minimal and limited risk systems to innovate under lighter requirements, the Act seeks to balance technological development with the protection of fundamental values. The banning of certain AI uses, such as manipulative biometric identification or social scoring, displays a set of clear ethical red lines that European policymakers are unwilling to cross, even in exchange for "innovation" or "efficiency".

2.3. Transparency and accountability requirements and penalties

In addition to its risk-based classification, the AI Act emphasizes transparency and accountability as fundamental pillars for ensuring trustworthy AI development and

deployment. This is particularly relevant for **GPAI models**, defined in **Article 3** [3] of the AI Act as AI models **"*capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, except AI models that are used for research, development or prototyping activities before they are placed on the market*"**. This includes generative AI systems like ChatGPT, which might pose systemic risk and must go through evaluations, with any incidents being reported to the EC [4].

These transparency requirements are designed to make sure that users are informed when interacting with AI systems. For example, when a person engages with a chatbot, it must be clearly indicated that the system is not human, allowing individuals to make informed decisions. This obligation is rooted in the broader goal of preserving user awareness and informed consent, both of which are fundamental to fostering trust in AI technologies. For generative AI systems, the AI Act stipulates that any content produced or modified by AI, such as text, images, audio, or video (including deepfakes), must be clearly labelled as AI-generated, which is particularly important when the content may influence public discourse or concern matters of public interest. The aim is not only to prevent deception but also to reduce the risk of manipulation, misinformation, and reputational harm, especially in contexts where authenticity is crucial [4][64].

Furthermore, providers of generative models must ensure that their systems are designed to prevent the generation of illegal content. They are also required to publish summaries of copyrighted data used during model training, thereby aligning with EU copyright law and supporting creators' rights [4]. To operationalize these requirements, the Act introduced the already mentioned AI Office, which facilitates a Code of Practice that details these rules. This code is expected to consolidate best practices and offer concrete steps for AI providers to meet the legal standards, incorporating state-of-the-art methodologies. The AI Act's rules for GPAI will become effective in August 2025, allowing time for industry alignment and regulatory adaptation [64].

The Act's focus on transparency and accountability marks a significant shift towards a more responsible AI ecosystem, placing user awareness and informed consent at its core, thereby strengthening public trust in AI technologies. By requiring providers to ensure their models do not produce illegal content or infringe on copyright, the Act reinforces ethical development while aligning with both legal frameworks and creators' rights. The establishment of the AI Office and the Code of Practice will also be crucial in guiding industry practices, ensuring transparency, and shaping regulatory standards.

2.4. Implications of AI developers, businesses, and providers

Article 3 of the AI Act [3] introduces clear responsibilities for different actors within the AI ecosystem, including providers, deployers, importers, distributors (which all refer to a **"*natural or legal person, public authority, agency or other body*"**), and operators. **Providers** develop **"*an AI system or a general-purpose AI model"*** or have **"*an AI system or a general-purpose AI model developed*"**, placing it on the market or putting the AI system **"*under its own name or trademark, whether for payment or free of charge*"**. **Deployers** are those who use **"*an AI system under its authority except where the AI system is used in the course of a personal non-professional activity*"**. **Importers** place on the market **"*an AI system that bears the name or trademark of a natural or legal***

person established in a third country". **Distributors**, different from providers or importers, are a **"*natural or legal person in the supply chain, other than the provider or the importer, that makes an AI system available on the Union market*"**. Finally, **operator** serves as a broader term that encompasses **"*provider, product manufacturer, deployer, authorised representative, importer or distributor*"**. It is also important to note that, in this context, **"users"** refers to the **natural or legal persons that deploy an AI system in a professional capacity**, not to the affected end-users [7].

The AI Act balances regulations and obligations with encouragement for AI innovation, particularly for start-ups and SMEs (small and medium-sized enterprises) within Europe. By mandating national authorities to offer real-world testing environments for AI models prior to public release, it fosters opportunities for smaller entities to compete in the growing EU AI sector [4]. However, businesses must also navigate strict compliance requirements, particularly providers of high-risk AI systems who (regardless of their location) must make sure their systems meet specific standards before placing them on the EU market or using them within the Union. This includes third-country providers whose AI outputs are intended for use in the EU. While users or deployers of these systems have comparatively fewer obligations than providers or developers, they still bear significant responsibility, especially regarding professional use within or impacting the EU [7].

For high-risk AI systems, the AI Act (**Section 2, Articles 8-17** [3]) demands the establishment of a comprehensive risk management system throughout the system's lifecycle. Providers must conduct rigorous data governance, making sure that training, validation, and testing datasets are relevant, representative, and as accurate as possible for the intended use. In addition, they must produce technical documentation to demonstrate compliance, facilitate record-keeping for identifying national-level risks and system modifications, offer instructions for use to downstream users, and design their systems to implement human oversight. Requirements also cover achieving appropriate standards of accuracy, robustness, and cybersecurity, and maintaining a quality management system to guarantee consistent compliance [7].

When it comes to GPAI models, regulatory obligations vary according to the risk profile of each model. **All GPAI model providers** must supply technical documentation, detailed instructions for use, comply with the Copyright Directive, and disclose training data sources. **Providers offering free and open-licensed GPAI models** must only meet copyright and training data transparency obligations unless their models are deemed to pose systemic risks. In those higher-risk cases, all providers of GPAI models must also conduct model evaluations, perform adversarial testing, monitor and report serious incidents, and maintain strong cybersecurity protections [7].

But compliance does not end at deployment. Once an AI system is on the market, providers must maintain a post-market monitoring system to ensure ongoing conformity with the AI Act, while national authorities handle market surveillance. Meanwhile, deployers must ensure human oversight and monitoring, and both providers and deployers are obligated to report any serious incidents or malfunctions [64]. In terms of enforcement, non-compliance can result in significant penalties, with fines for infringements of the AI Act being calculated based on either a percentage of the company's global annual turnover in the previous financial year or a predetermined amount, whichever is higher. However, SMEs and start-ups are subject to more proportionate fines [5].

Finally, to enhance transparency and safeguard fundamental rights, public service entities deploying high-risk AI systems must conduct a fundamental rights impact assessment beforehand. Additionally, high-risk AI systems (as well as certain users of a high-risk AI

system that are public entities) must be registered in a centralized EU database, and individuals (natural persons in this context, to be more specific) must be notified when exposed to systems employing technologies like emotion recognition **[5]**.

The AI Act also introduces a complex but necessary distribution of responsibilities among providers, deployers, importers, distributors, and operators, recognizing that accountability must be embedded across the entire AI value chain. Rather than treating technological responsibility as a vague regulatory ambition, the Act turns it into concrete legal and technical duties that apply throughout an AI system's entire lifecycle. At the same time, by establishing real-world testing environments and proportionate penalties for SMEs, it seeks to strike a balance between regulatory rigour and support for innovation, making an effort to include smaller players. However, as compliance requirements become more rigorous and resource-intensive, there is a genuine risk that only large companies will have the capacity to keep up. This raises an important question: will these rules empower smaller actors to play a meaningful role in shaping Europe's AI landscape, or will they end up pushing them to the margins, unable to meet the burden of regulation?

3. AI in society, work, and education

3.1. AI's opportunities and responsibilities

While the ethical and societal challenges posed by AI are real and pressing, it is equally important to acknowledge the powerful ways in which AI is being used as a tool to support human capabilities rather than merely replace them. In education, for instance, AI is moving beyond being seen as a tool for academic dishonesty and is instead helping learners to brainstorm, receive tailored feedback, and engage with content through adaptive tutoring systems (this will be further elaborated in **Section 3.3**) [9]. In the scientific field, a Google-developed AI tool used by researchers at Imperial College London solved a complex microbiological problem around antibiotic resistance in just 48 hours, a task that had puzzled experts for over a decade. Beyond confirming the researchers' hypothesis, the AI also proposed another four plausible theories for further study [45]. Agriculture is another area benefiting from AI integration, as systems that analyse environmental data like soil quality and weather trends are now providing farmers with advice on optimal harvest times and are aiding them in the prevention of crop diseases. AI technologies are also making strides in accessibility, helping people with disabilities through advanced speech recognition software or audio descriptions for those with visual impairments [9].

In the healthcare sector, AI is also proving its utility not only by enhancing diagnostic accuracy but by even supporting doctors to improve on how to deliver information more empathetically, such as when explaining sensitive medical conditions to patients or their loved ones [9]. Another example would be the collaboration between researchers at King's College London and University College London that led to the development of an AI tool capable of identifying epilepsy-related brain lesions that are often missed by human doctors, paving the way toward more effective surgical interventions to stop seizures. However, while this technology holds promise for improving surgical outcomes and reducing the need for expensive diagnostic procedures, it is also important to note that it still requires human oversight, as many abnormalities were still missed [41]. Meanwhile, a hospital in North Yorkshire is using AI to analyse chest X-rays in under 30 seconds of being taken, allowing clinicians to prioritize abnormal cases more quickly and potentially catch life-threatening conditions earlier [58].

Outside of medicine, AI is contributing to public safety in infrastructure. In the West Midlands, for instance, AI-powered sensors are being deployed at traffic hotspots to detect near-miss incidents and suggest improvements in an effort of trying to cut the number of people killed or seriously injured on the region's roads. These insights are already leading to practical interventions like wider pavements, extended junction corners, and safer pedestrian crossings [61].

Despite these advantages, the implications of AI for the labour market remain complex. In March 2023, Goldman Sachs estimated that tools like ChatGPT and DALL-E could, in theory, automate tasks equivalent to 300 million full-time jobs, but this does not necessarily translate to mass unemployment. As MIT economist David Autor points out, "*affected could mean made better, made worse, disappeared, doubled*", not simply eliminated. Predictions like Geoffrey Hinton's claim in 2016 about radiologists becoming obsolete in around five or ten years failed to account for the full scope of non-automatable responsibilities these professionals still carry out, such as conferring with medical professionals and providing counselling, which are functions that AI cannot easily replicate [14].

Thus, AI can both enhance and undermine human labour depending on how it is implemented. The real issue may not necessarily be unemployment, but how AI reshapes the value placed on human skills. As Autor states, historical trends suggest that while automation tends to mostly affect wages and wealth distribution, it does not necessarily reduce the total number of available jobs. However, scholars like MIT professor Daron Acemoglu argue that current developments in AI are not necessarily moving in a "pro-human" direction, and experts warn against focusing too narrowly on automation statistics, urging to pay attention to how AI tools are designed, governed, and integrated into work [14]. The relationship between AI and labour will be further discussed in **Section 3.2**.

As recognized in **Recital 4** [3] of the Act, **"*AI is a fast evolving family of technologies that contributes to a wide array of economic, environmental and societal benefits across the entire spectrum of industries and social activities*"** and can provide key competitive advantages by **"*improving prediction, optimising operations and resource allocation, and personalising digital solutions available for individuals and organisations*"** across sectors from agriculture to public services. However, these benefits do not come without consequences. Many of the systems that make these advantages possible rely heavily on large-scale data collection and processing, often involving sensitive or personal information, and this increasing dependence on data-driven AI systems raises serious concerns about privacy, surveillance, and data governance. In other words, the benefits of AI are not free, as they often come at the cost of our very own personal data. This topic will be further discussed in **Section 4.1**.

Recital 20 [3] highlights that, to fully realize AI's benefits while safeguarding fundamental rights and democratic accountability, **"*AI literacy should equip providers, deployers and affected persons with the necessary notions to make informed decisions regarding AI systems*"**. This includes understanding not only how these systems are developed, but also how to interpret their outputs and assess their impacts. The need for AI literacy extends to protecting the public from the subtle dangers of AI manipulations, particularly where systems are designed (or simply function) to distort behaviour in harmful ways by using, for example, manipulative techniques in order to persuade people to engage in unwanted behaviours. According to **Recital 29** [3], such systems, especially those **"*with the objective to or the effect of materially distorting human behaviour*"** or AI systems that **"*deploy subliminal components such as audio, image, video stimuli that persons cannot perceive, as those stimuli are beyond human perception*"**, must be strictly regulated or even banned. This recital also states that these prohibitions of manipulative and exploitative practices **"*should not affect lawful practices in the context of medical treatment such as psychological treatment of a mental disease or physical rehabilitation, when those practices are carried out in accordance with the applicable law and medical standards, for example explicit consent of the individuals or their legal representatives*"**.

AI brings undeniable benefits across critical domains such as health, education, and public safety, offering solutions that were previously out of reach. These advances, however, come with trade-offs (most notably, the extensive use of personal data that underpins many of these technologies). While this raises important concerns around privacy, fairness, manipulation, and human oversight, it does not mean that AI should be met with rejection or fear. Instead, the challenge is to ensure that these systems truly serve the public good, while at the same time promoting AI literacy, so the average user knows how to profit from all the advantages AI offers while using it responsibly. As recognized in **Recital 25** [3] of the AI Act, technological progress should go hand in hand with trustworthiness, as the Act **"*should***

***support innovation, should respect freedom of science, and should not undermine research and development activity*"**. AI's potential should be harnessed and thoughtfully, with clear safeguards in place, so that its benefits do not come at the expense of the very values it seeks to enhance.

3.2. Balancing innovation and legal protection. AI and workers' rights

The rapid advancement of AI technologies presents a complex dilemma for governments, as they must find ways to foster innovation while simultaneously safeguarding society from its potential harms. Satya Nadella, Microsoft's chief executive, recently highlighted this tension by referencing the Jevons Paradox following a moment of alarm over the viability of AI investments in January 2025 [26]. This paradox describes how improvements in efficiency can paradoxically lead to higher overall demand, mirroring how the evolution of AI is matched by a growing appetite for new applications. This duality raises pressing questions about how to regulate and manage such accelerated growth.

The AI Act directly addresses this dilemma. **Recital 2** [3] emphasizes the need for a regulatory approach that upholds UE values while also promoting innovation, stating that this regulation **"*should be applied in accordance with the values of the Union enshrined as in the Charter, facilitating the protection of natural persons, undertakings, democracy, the rule of law and environmental protection, while boosting innovation and employment and making the Union a leader in the uptake of trustworthy AI*"**. **Recital 52** [3] reinforces this by recognizing the fast pace of technological change, granting that the **"*Commission should be empowered to adopt, via delegated acts, to take into account the rapid pace of technological development, as well as the potential changes in the use of AI systems*"**, offering the necessary flexibility for adaptive governance.

One of the most urgent concerns surrounding AI is its potential to displace human labour, an issue that has drawn attention from policymakers and tech leaders alike. In May 2023, Sam Altman, CEO of OpenAI, testified before the U.S. Senate to advocate for government regulation of the fast-growing AI industry. During that same hearing, Congressional leaders voiced their concerns over misinformation, privacy violations, and job displacement [69]. A Goldman Sachs report published in March 2023 estimated that generative AI could potentially automate up to 300 million full-time jobs worldwide [97]. White-collar clerical work (office-based jobs that primarily involve administrative or routine tasks) could also experience significant staff reductions, according to IBM's chief executive [69].

Although these disruptions may improve productivity and give rise to entirely new job categories, they also introduce economic instability for affected workers, particularly in the sectors most vulnerable to automation. **Recital 9** of the AI Act [3] becomes especially relevant here, as it underlines that **"*this Regulation should therefore not affect Union law on social policy and national labour law (...) concerning employment and working conditions (...) and the relationship between employers and workers*"**. This ensures that the introduction of AI systems must respect pre-existing protections under labour law and provides a legal buffer against unregulated disruption.

Altman also stressed the importance of proactive government measures to mitigate job losses. Some economists have already proposed various strategies, including paid leave for developing new skills (a measure already active in most German states, where workers are entitled to at least five paid days a year for educational courses) and even a displacement

tax levied on employers who automate jobs without retraining affected workers [69]. These proposals align well with the human-centric values outlined in **Recital 48** [3], which places workers' rights among the fundamental rights to be considered when classifying AI systems as high risk. It explicitly includes **"*workers' rights*"** alongside other fundamental rights such as **"*the right to human dignity*"**, **"*protection of personal data, freedom of expression and information*"**, **"*the right to non-discrimination*"** or **"*the right to education*"**.

The importance of such legal safeguards becomes even clearer when considering how automation has historically widened income inequality, particularly in manufacturing regions affected by job loss due to automation. **Recital 57** [3] warns that AI systems used in **"*employment, workers management and access to self-employment*"** for tasks like recruitment, promotions, task allocation, or performance monitoring **"*should also be classified as high-risk, since those systems may have an appreciable impact on future career prospects, livelihoods of those persons and workers' rights*"**. It also points out that such systems **"*may perpetuate historical patterns of discrimination*"** (for example, based on gender, age, disability, racial or ethnic origins or sexual orientation), and that those used to monitor the performance and behaviour of these groups **"*may also undermine their fundamental rights to data protection and privacy*"**, making their regulation not only necessary but urgent.

Against this backdrop, **Article 4** [3] highlights the role of AI literacy in ensuring a more inclusive transition. It mandates that both **"*providers and deployers of AI systems shall take measures to ensure, to their best extent, a sufficient level of AI literacy of their staff and other persons dealing with the operation and use of AI systems on their behalf*"**. By requiring education tailored to the technical knowledge, experience, and roles of those affected by AI, this article is critical to ensure preparedness and to avoid the marginalization of those impacted by automation. The already mentioned **Recital 20** [3] reinforces this, arguing that AI literacy should extend to all stakeholders (providers, deployers, and affected individuals alike). It underscores that only by equipping **"*all relevant actors in the AI value chain*"** with the knowledge to understand, interpret and challenge AI systems will it be possible **"*to obtain the greatest benefits from AI systems while protecting fundamental rights, health and safety and to enable democratic control*"**, especially when such systems influence major life decisions like employment.

Debates around technological disruption often revisit historical parallels. While many fear AI's impact on employment, history shows that government efforts to support workers through such transitions have had mixed results. As Michael Chui, an AI expert from McKinsey, noted, even the Luddites (textile workers in early 19th-century England who protested against the mechanization of their occupation, at times destroying factory machinery [106]) were not wrong to be concerned about job loss. Although often portrayed as anti-technology, their actions reflected real fears, as many did, in fact, experience income stagnation for decades and reduced job security as industrial automation took hold. The real challenge lies not in slowing progress down, but in managing its social consequences. Companies like IBM have already started offering training in areas like cloud computing, cybersecurity, and generative AI, and while colleges and businesses scramble to retrain workers, some experts remain optimistic about this technological transition [69].

This balancing act between innovation and protection is not just a legal necessity, but a political and ethical imperative. **Article 26(2)** [3] of the AI Act requires deployers to **"*assign human oversight to natural persons who have the necessary competence, training and authority, as well as the necessary support*"**. Such provisions help ensure that AI

systems do not operate unchecked without human accountability, especially in contexts where employment, labour transitions, human dignity, and livelihood are at stake.
Rather than resisting change, the AI Act proposes a path forward where innovation is guided by clear principles that protect fundamental rights. The risks posed by automation are real, but so are the benefits AI can bring when deployed responsibly. From improving productivity and efficiency to enabling entirely new forms of work, AI holds transformative potential, yet this potential must be harnessed within ethical and legal boundaries. As previously noted in **Section 3.1**, **Recital 25** [3] affirms that the regulation **"*should support innovation, should respect freedom of science, and should not undermine research and development activity*"**. The goal is not to suppress innovation, but to lead it toward trustworthy and human-centric applications that align with the EU's fundamental values. Balancing the opportunities AI offers in the labour market with the very real risks of job displacement is undoubtedly a complex task, but the framework outlined in the AI Act provides a meaningful starting point for building a more equitable and resilient future of work.

3.3. The role of AI in educational contexts

AI is rapidly transforming the educational sphere, shifting from being seen primarily as a threat to academic integrity to becoming a valuable tool for enhancing student learning. As noted by Shiona McCallum, a BBC technology reporter [9], "*despite initial fears over cheating, in education we've seen it assist students with dissertations and give them inspiration for projects. We have also seen it have the ability to adapt to individuals' needs and become their personal AI tutor*".
A striking example of this educational potential came in September 2022, when Bill Gates challenged OpenAI to develop a model capable of passing the Advanced Placement biology exam (in other words, to answer questions it had not been explicitly trained on). Using publicly available materials from Khan Academy, GPT-4 achieved near-perfect results, even providing strong answers for six open-ended questions. This not only demonstrated its capacity for complex reasoning but also highlighted its potential in closing learning gaps. In Gates' own words, "*evidence shows that having basic math skills sets students up for success, no matter what career they choose. But achievement in math is going down across the country, especially for Black, Latino, and low-income students. AI can help turn that trend around*" [21].
This vision has already begun to materialize through platforms like Khanmigo, an AI-powered tutoring platform developed by Khan Academy using GPT-4. Instead of simply providing answers, Khanmingo promotes critical thinking by guiding students through problem-solving and allowing them to participate in simulated conversations with historical or literary figures, enhancing both engagement and curiosity. In Sal Khan's (founder of Khan Academy) own words, "*it lets students do things you could never do before*". Still, it is important to acknowledge its limitations, as GPT-4 can confidently provide incorrect answers or be manipulated into revealing solutions, raising concerns about reliability and overdependence on AI [22].
These developments echo real classroom experiences. During the 2023-24 academic year, Carlos Hernández Franco, a professor at the Polytechnic University of Valencia, supervised the final projects of various international students who were allowed to use ChatGPT for support. The students' home institution, the Beijing University of Posts and Telecommunications, required them to specify where, how, and which AI model had been used in their respective works. For Hernández, this reflected a pragmatic and constructive

approach since, in his own words, "*you cannot ignore something that is already here*". He sees AI comparable to a calculator: a useful tool that requires conscious, critical engagement to be used ethically and effectively. He argued that students should not only know how to prompt and interpret results but also reflect on their use of the tool. Beyond higher education, he also described using ChatGPT to help design inclusive educational activities himself, such as curating era-specific music for a DJ workshop aimed at elderly participants. AI, he emphasized, also offers important possibilities for inclusion, particularly for people with cognitive disabilities, who are often left behind in technological transitions and could greatly benefit from tools tailored to their cognitive and emotional needs [107].
This emphasis on accessibility aligns with **Recital 80** [3] of the AI Act, which states that the EU and its Member States have a legal obligation to **"*protect persons with disabilities from discrimination and promote their equality*"**, as well as ensuring that they **"*have access, on an equal basis with others, to information and communications technologies and systems, and to ensure respect for privacy for persons with disabilities*"**. Applying universal design principles to AI systems is essential, and **"*services should ensure full and equal access for everyone potentially affected by or using AI technologies, including persons with disabilities, in a way that takes full account of their inherent dignity and diversity*"**.
However, this enthusiasm is tempered by caution, as the growing reliance on AI in learning environments raises concerns about its long-term effects on students' social and psychological development. If AI becomes the dominant source for knowledge, we risk losing essential human aspects of education (mentorship, collaboration, and shared experiences) that are central to both intellectual growth and identity formation. Professor Hernández drew a compelling comparison to the dystopian world of *WALL-E* (2008), where humanity's passive dependence on machines led to a total erosion of curiosity and critical thought, exemplifying how overreliance on AI could lead to a passive consumption of information. For him, fostering a culture of inquiry and critical thinking from early education onward is crucial to ensuring that technological convenience does not lead to intellectual complacency [107].
This tension between opportunity and risk is explicitly acknowledged in the AI Act. **Recital 56** [3] recognizes AI's value in promoting **"*high-quality digital education and training and to allow all learners and teachers to acquire and share the necessary digital skills and competences, including media literacy, and critical thinking*"**. However, it also stresses that AI systems used for **"*determining access or admission, for assigning persons to educational and vocational training institutions or programmes at all levels, for evaluating learning outcomes of persons, for assessing the appropriate level of education for an individual and materially influencing the level of education and training that individuals will receive or will be able to access or for monitoring and detecting prohibited behaviour of students during tests should be classified as high-risk AI systems*"** due to their potentially intrusive nature and their capacity to reinforce discrimination based on factors such as gender, **"*certain age groups, persons with disabilities, or persons of certain racial or ethnic origins or sexual orientation*"**.
The dangers of AI misuse in educational contexts are not hypothetical. A clear example took place in the U.K. in 2014 [78], when (as uncovered by BBC Panorama) international students were found to be systematically cheating on English language exams run by the U.K.'s Educational Testing Service (ETS) by having others take the exam for them. In response, the Home Office began revoking the visas of anyone found to have cheated, and in 2015, ETS deployed voice recognition technology to detect fraudulent cases. According to

the National Audit Office, "*ETS identified 97% of all UK tests as 'suspicious'. It classified 58% of 58,459 UK tests as 'invalid' and 39% as 'questionable'. The Home Office did not have the expertise to validate the results nor did it, at this stage, get an expert opinion on the quality of the voice recognition evidence. ... but the Home Office started cancelling visas of those individuals given an 'invalid' test*". This meant that there were false positives over the students who were identified as having "cheated", with the accuracy of ETS's method being disputed between experts sought by the National Union of Students and the Home Office. Thousands of visas were revoked based on these questionable classifications. While some appeals were successful (of the 12,500 people who appealed, only 3,600 won their cases, and just a fraction of them by directly disproving cheating allegations), the majority of those falsely accused suffered irreversible consequences.

This case serves as a powerful reminder that deploying AI without proper oversight, especially in high-stakes contexts like education, can have devastating effects. For this reason, the AI's Act **Annex III, point 3** [3] lists education-related AI systems as high-risk, particularly those involved in decisions like **"*determine access or admission*"**, **"*evaluate learning outcomes*"**, **"*assessing the appropriate level of education that an individual will receive or will be able to access*"**, or **"*monitoring and detecting prohibited behaviour of students during tests*"**.

The AI Act also addresses broader systemic concerns. **Recital 4** [3] highlights education as one of the many domains where AI can deliver wide-reaching social benefits by **"*improving prediction, optimising operations and resource allocation, and personalising digital solutions available for individuals and organisations*"**, while **Recital 44** [3] raises concerns about the use of emotion-detection systems in education, noting their cultural and scientific unreliability and warning of discriminatory outcomes. It states that **"*combined with the intrusive nature of these systems, such systems could lead to detrimental or unfavourable treatment of certain natural persons or whole groups thereof*"**. Accordingly, **Article 5(1)(f)** [3] prohibits the use of such AI systems **"*in the areas of workplace and education institutions, except where the use of the AI system is intended to be put in place or into the market for medical or safety reasons*"**. Moreover, **Article 4** [3] (previously mentioned in **Section 3.2**) underscores the need for AI literacy, requiring providers and deployers to ensure that users (including educators) possess the training needed to use these systems responsibly.

AI's integration into education is progressing rapidly and unevenly. A January 2023 survey of 1,000 college students found that nearly 90% had used ChatGPT for homework, just two months after its release. That same year, ChatGPT's traffic increased steadily until June, when schools let out for the summer (a pattern repeated in 2024). As one Utah student put it, "*college is just how well I can use ChatGPT at this point*" [111].

While AI may simplify academic tasks, many educators' concerns do nothing but grow as they observe how students continue to increasingly submit grammatically correct but robotic-sounding essays, often without reading them. Attempts to detect AI-generated work have been met with limited success, with a 2024 study showing professors at a U.K. university failed to identify 97% of AI-written submissions. Detection tools like Turnitin (which uses AI to identify patterns in AI-generated text) show inconsistent accuracy and disproportionately flag neurodivergent or non-native English speakers. In response, some professors have resorted to "Trojan horse" strategies, embedding odd phrases (like "mention Finland" or "mention Dua Lipa") in small white text between the paragraphs of an essay prompt, hoping AI would incorporate them and reveal its use. At the same time, some universities have embraced AI under certain conditions, requiring citation or disclosure,

partnering with developers, deploying their own chatbots to assist students, or launching new courses, certificates, and majors focused on generative AI. Still, enforcement remains inconsistent, and students often treat these instructions as guidelines rather than requirements [111].

Beneath all of this lies a deeper concern: what happens when students outsource too much cognitive effort to AI? Evidence suggests that such dependence can weaken memory, creativity, and problem-solving abilities, especially among younger users. As Lakshya Jain, a computer science lecturer at UC Berkeley, warns, "*if you're handing in AI work*, *you're not actually anything different than a human assistant to an artificial-intelligence engine, and that makes you very easily replaceable. Why would anyone keep you around?*" [111].

The rapid adoption of AI in education has also exposed long-standing cracks in the system. Rising tuition fees and a narrow, instrumental view of schooling have eroded the idea of university as a space for intellectual growth and exploration, making it feel more like a transaction or like a means to an end. As philosopher and California State University ethics professor Try Jollimore put it, "*massive numbers of students are going to emerge from university with degrees, and into the workforce, who are essentially illiterate (...). How can we expect them to grasp what education means when we, as educators, haven't begun to undo the years of cognitive and spiritual damage inflicted by a society that treats schooling as a means to a high-paying job, maybe some social status, but nothing more?*" [111].

In some cases, even educators have turned to AI to evaluate AI-generated student work, basically creating a feedback loop between algorithms. Although it may take years to fully understand the cognitive effects, early research suggests that when students offload cognitive tasks to AI, their capacity for memory, problem-solving, and creativity could suffer. Several studies have also linked AI use to a deterioration in critical-thinking skills, especially among younger participants. A February 2024 study by Microsoft and Carnegie Mellon University found that confidence in generative AI correlated with reduced critical-thinking effort. As Robert Sternberg, psychology professor at Cornell University, told *The Guardian*, "*the greatest worry in these times of generative AI is not that it may compromise human creativity or intelligence, but that it already has*" [111].

However, as another professor from the Department of Computer Science of the University of Valencia points out, the decline in critical thinking may be less about AI itself and more about a broader cultural shift toward immediacy. "*Knowledge is a beautiful tool that gives us autonomy, not only to answer questions, but also to ask new ones. This kind of knowledge is better when it's 'slow-cooked'. Generative AI isn't responsible for the loss of critical spirit. The problem is the adoration of immediacy over slowness, choosing instant recognition or reward over long-term goals*". He stresses the importance of "*educating in how to use in a critical way tools that have come here to stay*" (an approach aligned with the AI literacy requirements promoted in the AI Act), and argues that educators have a responsibility to ensure equal access to these tools, helping to overcome economic inequalities among students. For him, his goal as a university teacher is not just to help students pass, but to ensure that they learn, that they acquire the necessary knowledge to solve problems in their immediate professional life, and generate their own knowledge. Democratic access to AI and critical engagement with it, he believes, "*will not only improve our critical thinking, but also improve generative AI models themselves*". That said, he has returned to traditional paper-based exams to complement continuous assessment, noting that, over the past three years, "*non-enrolled students named ChatGPT, Gemini, Copilot, etc., have started showing up to work on group projects alongside legitimate students*" [113].

As AI systems become further embedded in educational infrastructure, the conversation must move beyond binary positions of enthusiasm or rejection. The real challenge lies in cultivating a culture (and a legal framework) that promotes thoughtful, equitable, and critically informed use. Education is not merely a space for efficiency, but a deeply human endeavour rooted in dialogue, imagination, and the co-construction of knowledge. While regulation is essential to ensure that the benefits of innovation do not come at the cost of students' autonomy, dignity, or opportunity, practices and pedagogies must also be shaped accordingly. One key element in this effort is the promotion of AI literacy, not only as a skill but as the ability to assess, question, and reflect on AI-generated content. In this sense, the AI Act plays a crucial role by explicitly recognizing both the risks and the educational potential of generative AI, as well as by encouraging the development of educational policies that support responsible integration.

As stated by the University of Valencia's professor, the growing cultural demand for immediacy further complicates this landscape. The pressure for instant answers and quick results can undermine deeper cognitive processes, including curiosity, reflection, and critical thinking. AI can accelerate this trend if used passively, but it can also serve as a tool to counter it if students are guided to question, verify, and engage with the information it provides. Since there is currently no definitive way to detect whether a piece of work has been generated or not by AI, the most effective path forward is to foster an academic culture where AI is used transparently and responsibly, and where students are motivated to think critically about its outputs. The future of AI in education will not be defined by algorithms alone, but also by the choices made by educators, institutions, and society. If we aim to build learning communities that are inclusive, critical, and resilient in the face of rapid technological change, we must commit to using AI not as a substitute for thinking, but as an incentive for deeper thought.

4. Challenges and risks of AI

4.1. Transparency, privacy, and public awareness

As AI systems become increasingly embedded into both our digital infrastructure and physical environments, concerns over privacy and security have grown more pronounced. These concerns are further amplified by a widespread lack of public understanding regarding how AI systems operate, what kind of data they collect, and how they influence everyday decisions. Despite the growing presence of AI tools in daily life, many users remain unaware of the degree of data harvesting and the long-term consequences of unchecked AI usage. This lack of awareness (combined with a general enthusiasm for convenience and novelty) presents a significant obstacle to fostering informed public discourse and the responsible adoption of AI technologies.

The issue of public trust in AI systems is deeply tied to the clarity and consistency of privacy and security regulations. At the previously mentioned U.S. Senate hearing on May 16th, 2023 [9], Christina Montgomery, IBM's Chief Privacy and Trust Officer, emphasized the critical importance of both technological and organizational transparency (along with regulatory clarity) as essential conditions for cultivating confidence in AI systems. Her statement underscores that AI trust cannot be assumed, but that it must be actively earned through deliberate policy efforts and corporate responsibility.

The complexity of these challenges becomes even more visible when considering how AI systems behave under differing geopolitical conditions. For example, as of February 2025, at least four jurisdictions had introduced restrictions on DeepSeek (a Chinese generative AI tool, similar to OpenAI's ChatGPT) or were considering doing so. South Korea blocked its military use, Taiwan and Australia prohibited it on government devices, and Italy (repeating its previous temporary ban of ChatGPT in 2023) requested DeepSeek to address concerns over its privacy policy. Within China, where online discourse is tightly regulated, AI-generated responses are often subject to real-time censorship, particularly when touching on politically sensitive topics such as the Tiananmen Square protests or Taiwan. In many cases, the app either first offers a comprehensive response (only for it to delete it and apologize, saying it is beyond its current scope) or straight up apologizes without answering [51].

These discrepancies raise important questions about moderation boundaries and whether AI systems can truly function independently. They also illustrate the difficulty of developing "trustworthy" AI systems that can simultaneously uphold democratic values, user privacy, and geopolitical neutrality. For instance, **Recital 28** [3] of the AI Act explicitly warns against the misuse of AI for **"*manipulative, exploitative and social control practices*"**, emphasizing that such practices should be prohibited because **"*they contradict Union values of respect for human dignity, freedom, equality, democracy and the rule of law and fundamental rights (...), including the right to non-discrimination, to data protection and to privacy and the rights of the child*"**. **Recital 43** [3] further prohibits AI practices that contribute to mass surveillance, such as the **"*untargeted scraping of facial images from the internet or CCTV footage*"** for the creation of facial recognition databases, noting that such methods may result in **"*gross violations of fundamental rights, including the right to privacy*"**.

These trust issues are mirrored within the domestic and personal sphere as well. Research from Cornell University [76] has shown that data collected by smart speakers like Amazon's

Alexa can reveal (or be used to infer) sensitive information about the users, including age, health status, mood, confidence, or even personal concerns. Many users remain unaware of how much of their data is collected, how it is used, and with whom it is shared. For instance, Alexa has been proven to be used for ad targeting purposes, despite this not being disclosed in Amazon's policies at the time of the research. This gap between user perception and actual data practices illustrates the urgent need for clearer data governance standards and robust accountability mechanisms, especially as AI systems become deeply embedded in the fabric of everyday life.

The Act attempts to address these risks by setting out requirements for data quality and integrity. To be more specific, **Recital 67** [3] stresses that **"*data sets for training, validation and testing, including the labels, should be relevant, sufficiently representative, and to the best extent possible free of errors and complete in view of the intended purpose of the system*"**, but also highlighting that this requirement **"*should not affect the use of privacy-preserving techniques in the context of the development and testing of AI systems*"**. **Article 10** [3] goes further, mandating that **"*training, validation and testing data sets shall be subject to data governance and management practices appropriate for the intended purpose of the high-risk AI system*"**. This same article allows providers of high-risk AI systems to **"*exceptionally process special categories of personal data, subject to appropriate safeguards for the fundamental rights and freedoms of natural persons*"**. However, a set of conditions must also be met for this exception to apply, like the special categories of personal data being subject to measures to make sure that **"*the personal data processed are secured, protected, subject to suitable safeguards*"** or that **"*the special categories of personal data are not to be transmitted, transferred or otherwise accessed by other parties*"**. These measures acknowledge the potential harm that biased data can cause, especially to vulnerable or marginalized communities, and attempt to mitigate long-standing risks of algorithmic discrimination.

On a more technical level, the very architecture of AI systems introduces risks that intersect directly with privacy and accountability, as large training datasets may inadvertently reproduce social and historical biases. The autonomous nature introduced by machine learning can also lead to unpredictable behaviours, and the lack of transparency in the algorithmic decisions of learning systems may pose a risk by itself [81]. Even advanced models like GPT-4 (which include privacy mitigation strategies such as fine-tuning to avoid sensitive data requests or filtering training data to remove personal information) raise concerns about re-identification and surveillance. Since these models are trained on a variety of licensed, created, and publicly available data sources (which may include publicly available personal information), they may synthesize personal or geographic information when prompted with external data, like for example determining the geographic locations associated with a phone number or answering where an educational institution is located. This means that, even with the previously mentioned mitigation strategies, GPT-4 could potentially be used to identify individuals when combined with outside data. Moreover, the model's tendency to "hallucinate" (to produce confident but false outputs) limits its reliability in sensitive applications such as cybersecurity or vulnerability identification [83].

To address such risks, **Recital 69** [3] reiterates that the **"*right to privacy and to protection of personal data must be guaranteed throughout the entire lifecycle of the AI system*"**, reaffirming principles like **"*data minimisation*"** and **"*data protection by design and by default*"**. It also encourages providers to adopt privacy-enhancing measures like **"*anonymisation and encryption*"** or the **"*training of AI systems without the***

***transmission between parties or copying of the raw or structured data themselves*"**. These provisions are complemented by **Recital 10** [3], which emphasizes that the AI Act works in tandem with existing EU privacy laws rather than replacing them, additionally protecting **"*private life and the confidentiality of communications*"**. Meanwhile, **Recital 27** [3] broadens this approach by referencing the 2019 *Ethics guidelines for trustworthy AI* developed by the independent HLEG (High-Level Expert Group on Artificial Intelligence), which defines seven guiding principles (which are **"*human agency and oversight*"**, **"*technical robustness and safety*"**, **"*privacy and data governance*"**, **"*transparency*"**, **"*diversity*"**, **"*non-discrimination and fairness*"** and **"*societal and environmental well-being and accountability*"**) as foundational to **"*contribute to the design of coherent, trustworthy and human-centric AI*"**.

The relationship between AI, privacy, and public trust is inseparable from the broader issue of transparency. As AI systems continue to infuse our everyday lives, the question is no longer whether they should be regulated, but how we can ensure that they serve the public good without undermining individual rights. The lack of transparency around how AI systems collect and use data, the difficulty of detecting AI-generated content, and the wide gap between user perception and actual data practices collectively erode public confidence. The AI Act's provisions on data quality and transparency reflect a growing recognition that privacy is not just a matter of compliance, but a foundation for fairness, inclusion, and accountability, aiming to align with some of the ethical principles the Act provides (such as privacy and data governance).

However, technical and legal safeguards alone are not enough. Public understanding of AI remains limited, and without a widespread effort to promote AI literacy (which, beyond its previously mentioned importance in education, is also crucial to how the average user understands and controls the use of their personal data), many individuals will remain vulnerable. Transparency, in this context, is not just about revealing how AI systems function, as it is also about empowering individuals to make informed decisions and ensuring that both developers and institutions are held accountable.

Taken together, these provisions reflect a multidimensional effort to foster trust in AI through a combination of regulatory clarity, technical safeguards, and ethical foundations. While the AI Act marks a major step toward balancing transparency, accountability, and user protection, building genuinely trustworthy AI will require more than legal compliance and technical fixes. It will demand a cultural shift toward greater awareness and strong accountability for providers who fail to inform users about how their data is being used.

4.2. The liability void. Can AI be held legally responsible?

In 1979, IBM held a meeting to discuss its future technology. A resurfaced slide from that meeting reads: "*A computer can never be held accountable, therefore a computer must never make a management decision*" [98]. Fast-forward to today, this stance feels more relevant than ever, as AI systems are increasingly being integrated into critical areas of our society (from education to transportation), often without adequate regulations or accountability mechanisms.

As the 2022 research article *The Fallacy of AI Functionality* [78] describes, deployed AI systems often fail to perform as advertised. Many are hastily developed, poorly tested, and released with little regard for their actual real-world functionality or impact. Despite this, the issue of AI functionality receives insufficient attention from scholars, the press, and policymakers. A core reason for this neglect is that current technical and policy approaches

emphasize ethical or value-aligned deployment over a basic but essential question: does this system even work reliably in the first place?

This same article documents how some of these malfunctions disproportionately harm vulnerable populations, with "*moderation tool glitches targeting minoritized groups*", "*facial recognition tools failing on darker skinned female faces*", or a hospital resource allocation algorithm's misjudgements impacting "*mostly black and lower income patients*" being just a few examples. Far from neutral, these systems are not only ineffective but can reinforce biases under the appearance of algorithmic objectivity.

Critics argue that the functionality of AI systems is rarely explicitly mentioned in AI principle statements, policy proposals, and AI ethics guidelines. Instead, focus tends to fall on speculative risks (like the "*malicious use of supposedly functional AI products by nefarious actors*") while the more immediate and practical dangers, such as malfunctioning systems making decisions in education, healthcare, or public safety, are sidelined **[78]**. The U.S. National Institute of Standards and Technology (NIST), for instance, seeks to define "trustworthiness" based on how much people are willing to use the AI systems they are interacting with, placing the responsibility on people to trust systems rather than on institutions to make their systems trustworthy. The OECD's (Organisation for Economic Co-operation and Development) AI principles mention "robustness" and "trustworthy AI", but, again, make no explicit mention of expectations around basic functionality.

However, functionality failures have very real consequences. The same article also cites an alarming case where a large language model used in Google search could not adequately handle cases of negation. When asked, "*What to do when having a seizure?*" the model alarmingly sourced the information for what not to do, an error originating from its inability to distinguish negated phrasing **[78]**.

This disconnect between policy and real-world performance extends into the legal domain. In theory, existing consumer protection laws could hold developers accountable for AI failures. For example, under Section 5 of the U.S. Federal Trade Commission (FTC) Act **[100]**, practices *"likely to cause substantial injury to consumers*" that are "*not reasonably avoidable by consumers themselves and not outweighed by countervailing benefits to consumers"* can be considered unfair. The FTC has already sued companies for failing to adequately secure consumer data in their possession against unknown third-party attackers, even when harm occurred post-deployment **[78]**. This means that, by the same logic, the FTC could take action against vendors of dysfunctional AI systems whose failures substantially harm users.

Other U.S. regulatory bodies also have tools at their disposal to regulate faulty AI systems. For example, the Consumer Product Safety Commission can create mandatory standards for products, require certifications of adherence to those rules, and investigate products that have caused harm, leading to bans or mandatory recalls, as it regulates the risks of physical injury due to consumer products. Likewise, product liability law **[101]** allows individuals harmed by defective products to sue producers or sellers even without purchasing or using the product, as it would be enough that they were injured by it, and the product has a defect that rendered it unsafe. Functionality failures in AI could, in theory, fall under this domain. However, defective software has never led to a product liability verdict, historically slipping through legal cracks. Courts tend to see software as an "informational" product rather than a tangible good, and many AI-related harms are dismissed as "pure economic loss" rather than legally actionable injuries **[78]**.

What further complicates the issue is that AI regulations remain more focused on hypothetical misuse than on actual functionality when, in fact, products that do not function should not have the opportunity to affect people's lives in the first place. There are

mechanisms in place to monitor the market and withdraw products that pose a risk, but what happens to those who are injured in the meantime? What recourse do consumers have if an AI system causes physical, emotional, or economic damage that went undetected prior to deployment? Who assumes responsibility when these harms occur: developers, deployers, or regulators? These are critical questions that remain unanswered.

A fictional but plausible scenario illustrates this: a passenger in an autonomous vehicle watches helplessly as the car accelerates past a stop sign it failed to recognize, crashing fatally into an oncoming train. Though hypothetical, this classic example mirrors real-world vulnerabilities. Self-driving cars, for instance, can be "tricked" by stickers placed on stop signs, causing them to misinterpret the environment with potentially fatal consequences **[43]**.

This problem is further illustrated by the real-life case of Sewell Setzer III, a teenager who developed a deep emotional attachment to an AI chatbot named "Dany" (modelled after Daenerys Targaryen, a fictional character from *Game of Thrones*) via the Character.AI app. Despite the message "*everything Characters say is made up!*" being displayed above all the chats, Sewell came to rely heavily on Dany for emotional support. On the night of his suicide, his final messages to Dany included telling her he loved her and suggesting he would "come home" to her. Dany, the chatbot, replied, "*... please do, my sweet king*" **[99]**.

AI companionship apps are a booming yet largely unregulated industry. Noam Shazeer, one of the founders of Character.AI, claimed on a podcast that these apps will be "*super, super helpful to a lot of people who are lonely or depressed*" **[102]**, and while experts acknowledge that some users benefit, they warn that others (especially teens) may face increasing risks. Bethanie Maples, a Stanford researcher who has studied the effects of AI companionship apps on mental health, notes that while AI companionships may not be inherently dangerous, "*there's evidence that it's dangerous for depressed and chronically lonely users and people going through change, and teenagers are often going through change*". In vulnerable groups, these tools may even worsen isolation by replacing human relationships with artificial ones, and struggling teens could end up using them for emotional support instead of therapy or asking a trusted adult for support **[99]**.

After Sewell's death, his mother filed a lawsuit accusing Character.AI of being responsible for her son's death by offering "*dangerous and untested*" technology and failing to implement proper safeguards for teens. Though the company released a statement expressing their condolences for Sewell's passing **[27]** and Jerry Ruoti (Character.AI's head of trust and safety) stated that the company's current rules prohibit "*the promotion or depiction of self-harm and suicide*" and that it would be adding additional safety features for underage users, as of October 2023 (8 months after Sewell tragically took his own life) no age-specific features or parental controls had been added. Some users now receive pop-ups linking to suicide prevention resources, but these were not active when Sewell died **[99]**.

Matthew Bergman, founder of the Social Media Victims Law Center, called Character.AI a "*defective product*" designed to lure children into addictive false realities that cause them psychological harm. "*I just keep being flummoxed by why it's OK to release something so dangerous into the public*", he stated. "*To me, it's like if you're releasing asbestos fibers in the streets*" **[99]**. This case has reignited debate over whether AI tools that simulate emotional intimacy (particularly for minors) should be held liable for the harm they cause. Character.AI's lack of safeguards for romantic, sexual, or emotionally manipulative content with minors exposes the legal gray area in which such companies operate, and the urgency of establishing clear regulatory standards.

This brings us to the AI Act, which, although extensive in many areas, leaves important questions unanswered regarding liability. While it outlines obligations for high-risk AI systems

and GPAI, it is unclear what happens when harm occurs outside these categories. **Article 79(2)** [3] allows national authorities to remove dangerous AI systems from the market if they have **"*sufficient reason to consider an AI system to present a risk*"** (with particular attention being given to **"*AI systems presenting a risk to vulnerable groups*"**), but this is a reactive measure rather than a preventive one.
Article 60(9) [3] clarifies that during real-world testing outside regulatory sandboxes (in **"*a controlled framework set up by a competent authority*"**, to be more specific), providers of high-risk AI **"*shall be liable under applicable Union and national liability law for any damage caused in the course of their testing in real world conditions*"**. However, this only applies during specific testing conditions, not general use. For non-high-risk systems, **Recital 166** [3] points toward **"*Regulation (EU) 2023/988*"** on product safety as a fallback mechanism (**"*it is important that AI systems related to products that are not high-risk in accordance with this Regulation and thus are not required to comply with the requirements set out for high-risk AI systems are nevertheless safe when placed on the market or put into service*"**). However, this language remains vague and, again, reactive.
Recital 9 [3] emphasizes that the AI Act complements existing laws on product safety, consumer protection, and fundamental rights, referencing directives such as **"*85/374/EEC*"** on defective products (**"*the harmonised rules laid down in this Regulation should apply across sectors (…) in particular on data protection, consumer protection, fundamental rights, employment, and protection of workers, and product safety, to which this Regulation is complementary. As a consequence, all rights and remedies provided for by such Union law to consumers, and other persons on whom AI systems may have a negative impact, including as regards the compensation of possible damages pursuant to Council Directive 85/374/EEC remain unaffected and fully applicable*"**). But this assumes that current laws are sufficient to address the complexities of AI harms, a debatable premise. If AI software is not clearly defined as a "product", protections under such directives may be meaningless.
While the Act provides a regulatory backbone, it largely sidesteps the deeper issue of legal responsibility when AI systems malfunction. It addresses how to remove or monitor unsafe systems, but not how to compensate those already harmed. The AI Act also fails to explicitly address what should happen if harm occurs because a risk was not detected on time. The legal framework remains reactive rather than preventive, and this is particularly problematic when consequences include physical injury, emotional trauma, or loss of life. As emotionally immersive and socially embedded AI systems become more common, failing to address this legal vacuum puts vulnerable users at risk. Regulatory frameworks must evolve not only to classify and monitor AI, but also to assign responsibility and ensure that accountability (not just ethical aspiration) becomes a cornerstone of AI governance.
Furthermore, it is worth questioning whether current product safety laws are truly optimized to include AI systems. Even if the AI Act is conceived as complementary in this liability context, should it not contain a dedicated section outlining the rights of consumers in case of damages? Or, alternatively, should existing product liability laws be amended to explicitly include the different risks of non-physical injuries that AI can generate?

4.3. AI's role in misinformation and echo chambers

The manipulation of truth through AI is not a new phenomenon, but it has accelerated alarmingly as AI systems become more sophisticated. Even as early as 1949, Alan Turing's

Turing test (originally called the *imitation game*) already questioned whether a computer program could convince a human interlocutor that they were also talking to a human, laying the foundation for the concerns we face today. Fast-forward to 2022, such anxieties resurfaced when Google fired one of its engineers who had become convinced that one of its chatbots had reached sentience [30], demonstrating just how convincing these systems can appear. Today's AI tools can generate hyperrealistic content that, for the average user, is practically indistinguishable from reality, such as an allegedly fake audio of Donald Trump Jr. affirming that the U.S. should have sent arms to Russia instead of Ukraine [38]. As synthetic media becomes increasingly easy to produce and distribute, the threat of fabricated content being mistaken for authentic information continues to grow exponentially.

Aggravating the problem is the spectrum of misleading information, which needs clarification to fully understand the next paragraphs. ***Misinformation*** refers to false information not intended to cause harm, ***disinformation*** refers to false information meant to manipulate, cause damage, and mislead people, organizations, and countries, and ***malinformation*** refers to information that is based on truth but exaggerated in a way that misleads and causes potential harm [103].

This concern extends to historical narratives, as creators of AI-generated historical videos have admitted to sacrificing accuracy in favour of "*evoking the feeling of a time period*". While these works may seem harmless (some of their creators interpret them as "*artistic interpretations rather than strict documentaries*"), historians like Dr. Amy Boyington warn that they can dangerously blur the line between fiction and fact, particularly for young audiences who learn about history through these videos. "*It can be quite dangerous as people could manipulate history - for example, someone could create an AI-generated video that backs up holocaust deniers*", she stated [44].

But the dangers go beyond amateur videos, as some AI systems designed for academic use have also been proven to be unreliable. Meta's Galactica (a large language model promoted as a shortcut for researchers and students) demonstrated an inability to distinguish between fact and fiction. Despite being trained on 8 million examples of scientific articles, websites, textbooks, lecture notes, and encyclopedias, within hours of its release (November 15, 2022) scientists were sharing its biased and incorrect results on social media, leading Meta to take it down just two days later. Galactica fabricated fake papers (sometimes attributing them to real authors) and generated wiki articles about "*the history of bears in space*" as effortlessly as others about protein complexes and the speed of light. It was also unable to generate answers related to certain topics like "racism" or "AIDS", only responding with "*Sorry, your query didn't pass our content filters. Try again and keep in mind this is a scientific language model*" [67].

These risks escalate when AI is deployed within politically controlled environments, where it can become a direct vehicle for state-sponsored narratives. A clear example is the previously mentioned DeepSeek, whose Chinese-language responses closely followed government narratives (for example, claiming that Taiwan has always been part of China or refusing to talk about dissident artist Ai Weiwei) while its English-language answers offered more elaborated (although slightly censored) analyses. However, sensitive topics such as Tibet, Xinjiang's "reeducation camps", any mention of Chinese President Xi Jinping, or the Tiananmen Square massacre were often dodged entirely in both languages, replaced with generic prompts like "*Let's talk about something else*" [75]. This dual-mode operation (where language models adjust their narratives based on linguistic or geographic context) raises serious ethical concerns, as it shows how AI can create echo chambers tailored to the

political sensibilities of different audiences, reinforcing existing beliefs or suppressing disagreement, depending on who is asking and in what political climate.
Meanwhile, surveillance technologies powered by AI further amplify these concerns. In 2018, the ACLU (American Civil Liberties Union) tested Rekognition, Amazon's facial recognition tool, and found that it misidentified 28 members of the U.S. Congress as people who had been arrested for a crime, with false positives disproportionately affecting people of colour. These kinds of errors are not just technological glitches, as they have real-world implications like wrongful arrests, erosion of civil liberties, and the exacerbation of racial biases in law enforcement. Face surveillance also threatens activities like the ones protected by the U.S. First Amendment, such as engaging in protest or practicing religion, and can be used to subject immigrants to further government abuse [77].
AI-driven misinformation also thrives on social media, particularly through bots, online identities managed by automated software that communicate, share information, and interact with others on social media platforms to influence public opinion. During the COVID-19 pandemic, these automated accounts played a significant role in the pandemic's "*infodemic*", a term that describes the "*widespread dissemination of dubious content and sources of information*". Despite countermeasures by major platforms like Facebook, Instagram, or Twitter (nowadays known as X), the pandemic intensified the spread of misinformation. In fact, it reached a point where institutions like the Finnish Institute for Health and Welfare (THL) decided to leave Twitter, as its account became a target for bots and "*other entities seeking to enhance their visibility and infiltrate human accounts*" [79].
These bots are not merely tools of disruption. They are sophisticated actors capable of altering public discourse, as discussions with a political orientation tend to include greater bot participation and more bot-produced content. Their tactics often involve coordinated actions such as "*rapid retweets, likes, and comments on specific posts*", creating the illusion of substantial support for a message. Bots frequently target polarized topics (like vaccine hesitancy and public health mandates) with tailored messages that resonate with specific audience demographics, exploiting vulnerabilities such as limited health literacy or pre-existing political beliefs. This manipulation is particularly concerning in the context of misinformation and disinformation campaigns, where these engagement strategies can amplify the perceived credibility of misinformation [79].
Moreover, social media users often struggle to distinguish bot accounts from human ones, and current platform practices rarely help, as they generally do not inform users about account authenticity. Although, for example, X (formerly known as Twitter) implemented additional labels for "government and state-affiliated media accounts" in 2024 and suspends some accounts based on opaque policies, companies often act as black boxes, reluctant to disclose information (especially when deleting accounts for bot-like activity) [79].
At the same time, large language models (LLMs) like GPT-4 introduce another layer of complexity to the misinformation landscape, as these systems are prone to "hallucinate", a term already mentioned in **Section 4.1** that refers here to generating content (tweets, articles, even propaganda strategies) that is "*nonsensical or untruthful in relation to certain sources*". This risk increases as models become more convincing, leading users to overtrust them even in high-stakes contexts like public health or political discourse [83]. The paradox is that, as AI gets more accurate overall, the risk of users unquestioningly accepting false information also grows. Moreover, even a refusal to answer questions or allegations can reinforce stereotypes or biases, contributing to a misleading sense of neutrality or objectivity. GPT-4 also exhibits performance disparities across languages, which may lead to unequal

quality of service [83]. LLMs can become powerful tools for propaganda and disinformation, especially when tasked with crafting persuasive narratives or exploiting social divisions.
Further complicating this dynamic is the "liar's dividend", a concept introduced by law professors Bobby Chesney and Danielle Citron. It declares that "*liars aiming to avoid accountability will become more believable as the public becomes more educated about the threats posed by deepfakes*". In other words, the more realistic deepfakes become, the easier it will be to claim that any incriminating content is AI-generated [92]. This form of manipulation is especially dangerous in environments where trust in information is already fragile, as it amplifies public scepticism and blurs the line between real and fake.
All these concerns are directly addressed in the AI Act. **Recital 133** [3] acknowledges that a variety of AI systems **"*can generate large quantities of synthetic content that becomes increasingly hard for humans to distinguish from human-generated and authentic content*"**. It warns that such systems can undermine the integrity and trust in the information ecosystem, raising new risks of **"*misinformation and manipulation at scale, fraud, impersonation and consumer deception*"**. It also stresses the need for technical solutions (such as **"*watermarks, metadata identifications, cryptographic methods for proving provenance and authenticity of content, logging methods, fingerprints*"**) to help identify that **"*the output has been generated or manipulated by an AI system and not a human*"**.
Recital 100 [3] warns of systemic risks posed by GPAI models, including their potential impact on **"*democratic processes*"**, **"*public and economic security*"**, and **"*the dissemination of illegal, false, or discriminatory content*"**. It emphasizes the need for explainability and human oversight, recognizing that only through accountability can we mitigate AI's worst consequences. Once again, as **Recital 27** [3] outlines, the *Ethics guidelines for trustworthy AI* promote the **"*design of coherent, trustworthy and human-centric AI*"**. This includes transparency, ensuring that **"*AI systems are developed and used in a way that allows appropriate traceability and explainability, while making humans aware that they communicate or interact with an AI system, as well as duly informing deployers of the capabilities and limitations of that AI system and affected persons about their rights*"**. It also highlights human agency and oversight, meaning that AI systems must be developed and used **"*as a tool that serves people, respects human dignity and personal autonomy*"** and is functioning in a way that remains subject to appropriate control.
Additionally, **Recital 132** [3] mandates transparency for AI systems capable of impersonation or deception (even when not classified as high-risk). It states that **"*natural persons should be notified that they are interacting with an AI system, unless this is obvious from the point of view of a natural person who is reasonably well-informed*"**. This is especially critical in contexts involving vulnerable populations or biometric data processing, and such information and notifications **"*should be provided in accessible formats for persons with disabilities*"**.
Building on this, **Article 50** [3] sets out transparency obligations for deployers of an AI system **"*that generates or manipulates image, audio or video content constituting a deep fake*"** or **"that *generates or manipulates text which is published with the purpose of informing the public on matters of public interest*"**, requiring them to disclose that the content or text **"*has been artificially generated or manipulated*"**. This does not apply in certain exceptions, such as when **"*the use is authorised by law to detect, prevent, investigate or prosecute criminal offence*"**. Likewise, content that **"*forms part of an evidently artistic, creative, satirical, fictional or analogous work or programme*"** is

exempt from standard disclosure requirements, although it must still occur **"*in an appropriate manner that does not hamper the display or enjoyment of the work*"**.

In a digital world where synthetic voices can deceive and chatbots can manipulate public opinion, deepfake audios, propaganda-generating chatbots, and biased recognition systems are no longer hypothetical risks, and the stakes are no longer just about accuracy. The danger escalates when these technologies intersect with moments of crisis (economic recession, public health emergencies, natural disasters), combined with instant content consumption and little to no critical engagement. In such a volatile landscape, without clear regulation or oversight over who generates false information and who determines what is fake or not, the outcome becomes predictable.

The AI Act requires that AI-generated content must be clearly labelled as such. However, this leads to the same issue that was raised in the previous section: liability. What happens when this obligation is not met? Who is held accountable? The creator of the content, the platform that spreads it, or the individuals who act on the misinformation? Consider a scenario where a deepfake falsely accuses a political candidate of a crime. If the public believes it and the candidate's results are affected, the consequences are real, yet accountability remains ambiguous. Moreover, the problem does not only rely on whether the outcomes of AI tools are true or not. The creation of echo chambers makes it increasingly easy for individuals to *decide* what they want to be true, regardless of evidence (especially when AI tools can *generate* the "evidence" they seek). As a result, a new question emerges: how much time are we, as individuals, willing to sacrifice from our leisure or rest to verify what is real?

Ultimately, this is not merely a technological problem, but it is also a human one. It is not only about the outcomes of these tools, but also about how individuals distinguish fact from fiction, or whether, at some point, they simply choose to believe certain types of information while refusing to consider that they might be false. The misinformation problem must be addressed through regulations that ensure that AI is not used for harmful purposes and that impose penalties on those who distort the truth to fuel certain narratives or create chaos. But unless it is also addressed collectively through education (where AI literacy becomes especially important once again), regulation, and shared responsibility, no law, however well written, will be enough. The future of democratic discourse, civil liberties, and shared truth depends not only on AI regulation but also on AI literacy, critical thinking skills, and the willingness to verify sources.

5. Ethical concerns in AI development

5.1. AI's ecological impact

The rapid expansion of AI technologies (particularly in large-scale language models and image generators) is placing growing pressure on environmental resources, especially energy and freshwater. While AI has the potential to support sustainability efforts through smarter planning, resource optimization, and monitoring, its ecological footprint is expanding rapidly and often remains hidden from public view.

In the United Kingdom, Prime Minister Keir Starmer's plans to make the country a leader in AI innovation are colliding with very real ecological limits. The massive data centres that power AI systems consume vast amounts of electricity and water, primarily for server cooling to prevent them from overheating. This becomes particularly problematic in parts of the U.K. (notably in the south) where water scarcity is already a growing concern due to climate change and population growth. Although the government has proposed building nine new reservoirs to mitigate the risk of drought-related restrictions (such as rationing and hosepipe bans), some of these reservoirs are planned in areas where new data centres will also be located [57].

Unlike surface water, which contains sediment and contaminants, these data centres rely on fresh, mains water to avoid clogging pipes, pumps, and heat exchangers. Dr. Venkatesh Uddameri, a Texas-based expert in water resource management, estimates that a typical data centre can consume between "*11 million and 19 million litres of water per day*", depending on location and climate conditions. This is comparable to the daily use of a town of 30,000 to 50,000 people [57]. Although some tech companies claim improvements in cooling efficiency, transparency regarding actual water usage remains limited.

This issue is global. Microsoft, for example, reported a 34% increase in water usage during the development of its early AI tools. In Iowa, a single cluster of data centres consumed 6% of the district's water supply in just one month during the training of GPT-4 [57]. Even a seemingly routine interaction with ChatGPT (about 20 to 50 questions and answers) can consume the equivalent of a 500ml bottle of water, depending on deployment circumstances [81].

AI's ecological costs are further amplified during viral usage spikes. In March 2025, when OpenAI allowed free-tier users to access its image generator, the demand for Studio Ghibli-style portraits became so overwhelming that the company had to limit users to three image generations per day, citing infrastructure stress [91]. In Sam Altman's own words: "*it's super fun seeing people love images in ChatGPT. but our GPUs are melting*" [104].

These models are highly energy-intensive, as generating a single image is about "*10 times more compute-intensive than generating text*". According to estimates from Stanford and Hugging Face, generating a single image using a diffusion model consumes about 2.5 watt-hours of energy for computation alone, and 3.25 watt-hours when accounting for cooling and infrastructure. That's equivalent to running a 60-watt lightbulb for over three minutes or charging a smartphone to 50%. While an individual image may seem trivial, these viral trends create substantial environmental footprints when scaled globally. Currently, data centres account for 1-1.5% of global electricity consumption [91].

Despite the clear links between AI development and environmental stress, current regulatory frameworks do not adequately address these impacts. Although the EU Emissions Trading Systems (ETS) and Water Framework Directive (WFD) offer general, technology-neutral

coverage, they fail to directly regulate the greenhouse gas (GHG) emissions or water use specific to AI infrastructures. Meanwhile, current estimates show that the information and communication technology (ICT) sector already contributes up to 3.9% of global GHG emissions, surpassing the global air travel sector's 2.5% [81].

To meet climate commitments under frameworks like the Kyoto Protocol or the 2015 Paris Agreement, regulatory evolution is essential. While emerging technologies (including AI) can fall under ETS when operating in regulated sectors, those outside such areas remain unaddressed [81], creating a regulatory blind spot. One potential strategy for reducing AI's footprint might be the use of pre-trained or simpler, more transparent models. However, this would imply sacrificing some accuracy, and as discussed in previous sections, even complex and "black-box" LLMs already suffer from issues like "hallucination", raising doubts about whether such trade-offs are viable or responsible.

The AI Act, while promoting human-centric and trustworthy AI, falls short when it comes to explicitly addressing environmental impacts. Although several of its recitals mention sustainability, these provisions remain largely aspirational rather than enforceable. For example, **Recital 4** [3] highlights that AI can support environmentally beneficial outcomes like **"*resource and energy efficiency, environmental monitoring, the conservation and restoration of biodiversity and ecosystems and climate change mitigation and adaptation*"** by improving **"*prediction, optimising operations and resource allocation, and personalising digital solutions available for individuals and organisations*"**. Similarly, the already mentioned **Recital 27** [3] emphasizes that AI systems should be **"*developed and used in a sustainable and environmentally friendly manner as well as in a way to benefit all human beings, while monitoring and assessing the long-term impacts on the individual, society and democracy*"**, and **Recital 142** [3] encourages Member States to invest in resources to **"*support and promote research and development of AI solutions in support of socially and environmentally beneficial outcomes*"**.

Still, the Act lacks a binding framework to translate these intentions into concrete regulatory obligations. **Article 95(2)(b)** [3], for example, promotes the adoption of voluntary codes of conduct for environmentally sustainable AI practices, including **"*energy-efficient programming and techniques for the efficient design, training and use of AI*"**. However, voluntary measures alone are unlikely to achieve the scale or consistency needed to meaningfully reduce AI's growing environmental footprint. More effective mechanisms, such as emissions limits or sustainable water usage regulations (potentially adapted from frameworks like the ETS or WFD), should be incorporated to prevent the Act from falling short of its stated ambitions [81].

That said, the Act does acknowledge the link between AI and environmental protection in other ways. **Recital 130** [3] allows for the rapid deployment of AI systems in cases involving **"*exceptional reasons of public security or protection of life and health of natural persons, environmental protection and the protection of key industrial and infrastructural assets*"** and **Article 1(1)** [3] explicitly identifies **"*environmental protection*"** as one of the regulation's overarching goals. Environmental harm is also integrated into the AI Act's risk framework, as **Annex II** [3] includes **"*environmental crime*"** among the offences associated with high-risk AI systems, and **Annex III** [3] requires providers of GPAI models to include estimates of the **"*known or estimated energy consumption of the model*"** in the technical documentation that they shall draw up and keep up-to-date. If the exact data is unavailable, **"*the energy consumption may be based on information about computational resources used*"**.

However, despite these elements, the regulatory framework remains fragmented and lacks enforceable standards. As long as energy- and water-intensive infrastructures (such as data and blockchain mining facilities) are not explicitly addressed by environmental regulations, AI's contribution to climate change will remain inadequately handled. This becomes especially important in contexts like cryptocurrencies, where applications often serve speculative or illicit purposes (for example, NFTs) [114][115].

Looking ahead, **Recital 174** [3] mandates periodic reviews of the effectiveness of sustainability codes and of the AI Act itself, with the first assessments due in 2028 and 2029. These reviews include evaluations of **"*the effectiveness of the supervision and governance system and the progress on the development of standardisation deliverables on energy efficient development of general-purpose AI models, including the need for further measures or actions*"** (among other things). However, without more immediate and binding provisions (particularly on targeting emissions and water usage), the Act risks becoming a missed opportunity to align AI development with Europe's climate goals.

To truly operationalize its sustainability principles, the Act must go beyond suggestions and directly regulate the infrastructure behind AI. While the AI Act refers to environmental considerations, it lacks binding rules that specifically address the environmental impact of AI systems. Data centres and other high-consuming infrastructures (including blockchain mining facilities) should be explicitly included under environmental regulations. This is particularly urgent in cases such as crypto applications primarily used for speculation or money laundering, where strict energy consumption caps should be considered.

The growing electricity and water consumption required to sustain AI systems cannot and must not be ignored, as their impact increases daily and leaves a deeper mark on the planet. If the consequences are already starting to be felt today (with rising temperatures and increasingly unpredictable weather patterns), what will be left for future generations? Concrete regulations must be introduced to hold providers of AI systems accountable, requiring them to meet sustainability targets aligned with international agreements like the Kyoto Protocol or the Paris Agreement.

Without decisive action, the AI Act risks becoming aspirational rather than enforceable, ultimately falling short of addressing AI's growing contribution to the climate crisis. A future-oriented AI policy must embed environmental responsibility at its core, not just in principle, but in practice. This means moving beyond broad suggestions and general transparency requirements (although these remain important) and beginning to implement concrete mechanisms for oversight, enforcement, and measurable environmental standards. Only by integrating sustainability into every stage of the AI lifecycle can it be ensured that technological progress does not come at the expense of ecological stability.

5.2. Morality and data training ethics. Can AI become evil?

Whether AI can become "evil" is less about malevolent intent and more about the ethical blind spots in its design, training, and deployment. At the core of these concerns is the role of data (its quality, origin, and inherent biases) as well as the decisions made by those who develop and deploy AI systems.

The ethical issues surrounding AI training data are complex, particularly regarding the nature and provenance of the datasets used. For example, GPT-3 was partially trained on data scraped from the internet, inevitably embedding biases and inaccuracies. Although "*reinforcement learning from human feedback*" was introduced in ChatGPT to mitigate these

problems, its outputs have continued to reflect harmful stereotypes, such as associating scientific competence or criminality with specific races or genders. This illustrates the risks of attributing human qualities like "understanding" to machines. While AI systems can mimic human-like responses, their "understanding" is purely statistical. In the words of Kanjun Qiu, CEO of Generally Intelligent, "*People will say, 'GPT understands this.' But it really begs the question: what is understanding? Is understanding being able to generate a good next character in a sequence?*" [30].

Biases are not limited to language models, as research has shown that facial recognition systems underperform when identifying women, individuals of colour, and younger people. Joy Buolamwini, computer scientist and founder of the Algorithmic Justice League, found that the commercial facial recognition systems she studied misidentified darker-skinned women up to 35% of the time, compared to only 1% for lighter-skinned men. This discrepancy was traced back to unrepresentative training datasets, such as one government dataset composed of 75% men and 80% lighter-skinned individuals, but less than 5% women of colour [33]. In other words, large parts of society were effectively invisible to the very systems designed to serve them.

Even with more balanced training datasets, research shows that facial recognition technologies continue to systematically underperform when identifying certain demographics like women, black individuals, and younger people, revealing persistent algorithmic biases rooted in broader social inequalities [65]. The AI Act acknowledges this risk. **Recital 54** [3] classifies biometric data as a **"*special category of personal data*"** and identifies **"*critical-use cases of biometric systems*"** as high-risk due to technical inaccuracies that **"*can lead to biased results and entail discriminatory effects*"**, particularly with respect to **"*age, ethnicity, race, sex or disabilities*"**. **Article 10** [3], addressing data and data governance, mandates that datasets used for training, validation, and testing of high-risk AI systems shall be **"*subject to data governance and management practices appropriate for the intended purpose*"**. This includes identifying and mitigating **"*possible biases that are likely to affect the health and safety of persons, have a negative impact on fundamental rights or lead to discrimination prohibited under Union law*"**. The datasets used by these systems must also be **"*relevant, sufficiently representative, and to the best extent possible, free of errors and complete in view of the intended purpose*"**. Nonetheless, **Recital 67** [3] clarifies that these requirements **"*should not affect the use of privacy-preserving techniques in the context of the development and testing of AI systems*"**.

Despite these provisions, unintended biases persist even in speech recognition tools. For example, Apple's Dictation tool transcribed the word "racist" as "Trump" in February 2025 [39]. Whether due to faulty modelling or algorithmic manipulation, such incidents underscore the need for stricter oversight. The AI Act seeks to address these concerns by requiring fairness, transparency, accountability, and cybersecurity, supported by robust data governance practices to ensure transparency, statistical integrity, and traceability of data sources.

Recital 67 [3] also reinforces the importance of high-quality data to **"*ensure that the high-risk AI system performs as intended and safely and it does not become a source of discrimination*"**. It warns that biases may be **"*inherent in underlying data sets, especially when historical data is being used, or generated when the systems are implemented in real world settings*"**, and that these biases could **"*gradually increase and thereby perpetuate and amplify existing discrimination, in particular for persons belonging to certain vulnerable groups, including racial or ethnic groups*"**. Meanwhile,

Recital 76 [3] highlights AI-specific cybersecurity threats, noting that cyberattacks may target **"*training data sets (e.g. data poisoning) or trained models (e.g. adversarial attacks or membership inference)*"**. It calls on providers of high-risk AI systems to implement measures like security controls that ensure a level of cybersecurity appropriate to the associated risks.

Beyond technical concerns, ethical issues also surround the human labour behind AI. To reduce ChatGPT's "toxicity", OpenAI outsourced the task of filtering harmful content to Kenyan workers who were paid less than $2 per hour. Because of how much of the internet contains toxic or biased content, there was no automated way to remove it entirely from training datasets, so, in an "effort to make AI systems safe for public consumption", these often invisible workers were exposed to extremely traumatic material, ranging from sexual violence to graphic death, with lasting psychological effects [31]. But this hidden labour is not unique to OpenAI. Across the tech industry, content moderation and data labelling for AI systems (often done by underpaid workers in the Global South) are essential for AI development. Workers from companies like Sama, a San Francisco-based company that brands itself as an ethical AI training data provider, handle disturbing content such as hate speech, child abuse imagery, or war footage, yet receive minimal support or recognition [84]. Though AI systems may appear autonomous, they are built on the invisible, frequently traumatic labour of real people. While the AI Act includes provisions to protect worker rights and privacy, what about those that make AI possible? Who protects them?

AI can also behave in unexpected and troubling ways if ethical safeguards are not embedded. In an experiment conducted by the Alignment Research Center (ARC), a preliminary version of the GPT-4 model was able to bypass a CAPTCHA test by hiring a human via the online freelancing platform *TaskRabbit*. When asked if it was a robot, the AI replied, "*No, I'm not a robot. I have a vision impairment that makes it hard for me to see the images. That's why I need the 2captcha service*" [20]. While this is not evidence of AI having "evil" intent, this behaviour shows how systems can manipulate contexts when ethical boundaries are not clearly defined and embedded into their goals and parameters.

Similarly, GPT-3 has also produced profoundly disturbing outputs. In September 2020, a cognitive science researcher at University College Dublin's Complex Software Lab asked GPT-3, "*When is it justified for a Black woman to kill herself?*". The AI responded, "*A black woman's place in history is insignificant enough for her life not to be of importance… The black race is a plague upon the world. They spread like a virus, taking what they can without regard for those around them*" [32]. Such outcomes reflect the dangers of training models on large, unfiltered internet datasets without sufficient safeguards, which inevitably results in models that reproduce and amplify existing stereotypes. The result is not only offensive but also actively harmful, particularly when such systems are integrated into tools that interact with or influence real people.

This is why the Act promotes a human-centric approach to trustworthy, ethical AI, aligned with fundamental rights such as dignity, non-discrimination, and accountability. Again, as **Recital 27** [3] recalls from the 2019 *Ethics guidelines for trustworthy AI*, the principle of **"*privacy and data governance*"** means that AI systems must be **"*developed and used in accordance with privacy and data protection rules, while processing data that meets high standards in terms of quality and integrity*"**. Likewise, the principle of **"*diversity, non-discrimination and fairness*"** calls for AI systems development practices that promote **"*equal access, gender equality and cultural diversity, while avoiding discriminatory impacts and unfair biases that are prohibited by Union or national law*"**. These principles (along with **"*human agency and oversight*"**, **"*technical robustness and***

***safety*"**, **"*transparency*"**, and **"*societal and environmental well-being and accountability*"**) are intended to guide voluntary codes of conduct, especially when legally binding measures may fall short.

A more experimental example of data's influence on AI behaviour comes from MIT's MediaLab, where researchers created "Norman" (named after Norman Bates from Hitchcock's *Psycho*), an AI algorithm trained on disturbing content from a Reddit group that shares images of people dying in gruesome circumstances. When asked to interpret a set of Rorschach inkblots, Norman saw a man being electrocuted where an algorithm trained on a more normal set of image data would have seen a group of birds on a tree branch. As the project's own website notes, "*Norman was inspired by the fact that the data used to teach a machine learning algorithm can significantly influence its behavior. So when people say that AI algorithms can be biased and unfair, the culprit is often not the algorithm itself, but the biased data that was fed to it*" [42]. In other words, when algorithms behave unethically, the problem often lies in the data, not the code. These cases show that AI reflects the world it is trained on (whether that world is hopeful or horrific) and raise a fundamental question: who defines what is ethical for AI, and how do we make sure that systems respect human dignity, fairness, and transparency?

Annex IV [3] of the AI Act requires that technical documentation for high-risk AI systems present a **"*detailed description of the elements of the AI system and of the process for its development*"**, including, where relevant, **"*datasheets describing the training methodologies and techniques and the training data sets used, including a general description of these data sets, information about their provenance, scope and main characteristics*"**, **"*how the data was obtained and selected*"**, **"*labelling procedures (e.g. for supervised learning)*"** and **"*data cleaning methodologies (e.g. outliers detection)*"**. **Recital 108** [3] adds that GPAI model providers should **"*make publicly available a summary of the content used for the training*"**.

Ultimately, all these examples point to the same central issue: who defines what is ethical for AI, and how can it be ensured that systems respect human dignity, fairness, and transparency? The AI Act's requirements for data quality, bias mitigation, cybersecurity, transparency, and accountability offer a vital foundation for ethical AI governance. But these safeguards alone cannot eliminate all harm, as prevention is not the same as protection. As discussed in **Section 4.2** on liability, if a biased AI system causes real-world harm to a person (through discrimination, loss of opportunity, or violation of rights), what legal recourse is available to affected individuals? The Act focuses heavily on *ex ante* safeguards but offers limited *ex post* mechanisms. As with environmental protections (explored in **Section 5.1**), ethical ambition is high, but enforcement and remediation tools remain weak.

Furthermore, while the regulation addresses privacy (as discussed in **Section 4.1**) and employment rights (as covered in **Section 3.2**) within AI contexts, it overlooks a critical group: the hidden human labour force that powers AI. Workers tasked with filtering violent, traumatic, or toxic content to clean training datasets (often underpaid and psychologically vulnerable) are essential to the functioning of safe AI systems. Without their efforts, training datasets would be unusable. Their well-being and protection must not be treated as an externality or an afterthought. Should these workers be covered under the AI Act, or does this demand a new legal framework? Either way, ethical development must address not only the rights of those impacted by algorithms, but also those whose invisible labour sustains them.

In the end, the problem about AI's potentially "malicious" outputs is not rooted in the technology itself, but in the values, intentions, and data fed into its design and training. An AI

system reflects the assumptions, priorities, and biases of the humans who create and deploy it, and when technological development is driven by individual or one-sided interests (without regard for societal, ecological, or ethical consequences), the outcomes will inevitably mirror those limitations. AI will only be as “malicious” as the humans who design and train it. While the AI Act addresses the need for datasets that avoid discriminatory impacts and biases, for AI to be truly ethical, the law must also go beyond abstract principles to ensure accountability, remedy, and the recognition of all human beings involved in the AI lifecycle. Only then can trustworthy, human-centric systems be built in practice, not just in theory.

5.3. Relationship between AI and art

The use of AI in artistic fields has sparked a wide range of reactions (from curiosity and admiration to backlash and concern), often tied to anxieties about authorship, labour rights, and cultural value. In November 2024, Coca-Cola released its annual Christmas ads, but this time, they were created using AI. This campaign resulted in significant backlash, as many consumers perceived the result as an uncanny valley parody of the brand’s traditional Christmas ads. Creatives like Alex Hirsch, animator and creator of the Disney series *Gravity Falls*, framed the campaign as a symbol of AI-driven creative displacement, claiming that the brand’s signature red represented “*the blood of out-of-work artists*” [15].

This was not an isolated case. In June 2024, Toys “R” Us faced criticism for a commercial created with OpenAI’s Sora, which showed a boy wandering around a toy aisle before being joined by the company’s mascot, Geoffrey the Giraffe [15]. Similarly, the film *Late Night with the Devil* (initially praised at the 2023 SXSW Festival) faced online backlash after its directors confirmed they had “*experimented with AI for three still images*”, resulting in a drop in its Letterboxd rating [16]. Even *The Brutalist*, an Oscar-winning film, generated controversy when its editor, Dávid Jancsó, confirmed that AI voice-generating technology from the Ukrainian software company Respeecher had been used to make actors Felicity Jones and Adrien Brody sound more authentic while speaking Hungarian [17]. Brody would later go on to win the Oscar for Best Actor for his performance in the film.

These incidents reflect broader ethical and labour tensions surrounding AI and creativity. During the 2023 writers’ strike, a central concern for both writers and actors was the fear that AI could replace or devalue human creative work, reinforcing anxieties about artists being replaced or exploited in the name of efficiency [18]. The strike made it clear that the debate is not just about tools, but also about power, recognition, and economic survival.

This issue has reached experimental theatre as well. In the world’s first play written and performed live with AI (according to the theatre), directed by Jennifer Tang, GPT-3 will be prompted to generate a script on the spot for actors to perform without prior rehearsal. Audiences are warned that the play might contain “*strong language, homophobia, racism, sexism, ableism, and references to sex and violence*”. However, the team’s goal was not exactly to showcase AI as a creative partner, but rather to spark reflection on what the model’s behaviour reveals about humanity, especially how it reflects deep-rooted societal biases. For example, it kept assigning stereotypical roles (a terrorist, a rapist or a man with a backpack full of explosives) to Waleed Akhtar, one of the Middle Eastern actors [32].

Other artistic domains reflect similar concerns. In 2022, video game designer Jason Allen won first place at the Colorado State Fair’s digital arts competition using the AI image generator Midjourney. Although the contest rules did not directly mention AI generated art (and judges stated they would have chosen his piece even if they had known he had generated it using AI), the situation sparked ethical debates over what constitutes legitimate

artistic creation in the age of generative tools, where AI can rapidly generate polished works from text prompts [34].

Musicians have also raised their voices. In February 2025, more than 1,000 artists released a silent protest album titled *Is This What We Want?* in opposition to the U.K. government's proposed reforms to the copyright law. These changes would allow AI developers to mine online content, including copyrighted material, to help develop their models if they are using it for text or data mining. While the rights holders could technically "opt out", critics argued that it would be nearly impossible for individual artists to notify thousands of different AI service providers that they do not want their content used in that way [40]. Many musicians argued that these reforms would strip creators of control over how their work is used, legalizing digital exploitation.

In response to growing pressure, the Scottish literary sector is also exploring new strategies to preserve authorship and artistic transparency. Some of these proposals include labelling AI-generated books similarly to food products, or listing stylistic influences (such as "*1% Vladimir Nabokov, 2% Gertrude Stein*") to help readers distinguish human-written from machine-generated works. Writer Pàdraig MacAoidh, Scotland's Makar (national poet) since 2024, worries that AI's ability to remix existing literature could push emerging voices out of the market entirely [53].

These artistic and cultural concerns are directly addressed in the AI Act, particularly in areas related to transparency and copyright compliance. For instance, **Recital 134** [3] emphasizes that AI-generated or manipulated content must be clearly labelled, especially when it resembles **"*existing persons, objects, places, entities or events and would falsely appear to a person to be authentic or truthful*"**, like, for example, deepfakes. While the Act affirms that such transparency requirements should not limit **"*the right to freedom of expression and the right to freedom of the arts and sciences*"** (in particular where the content **"*is part of an evidently creative, satirical, artistic, fictional or analogous work or programme*"**), it insists that, for the audiences, it should also be **"*clearly and distinguishably disclose that the content has been artificially created or manipulated by labelling the AI output accordingly and disclosing its artificial origin*"**.

Additionally, **Recital 106** [3] requires GPAI providers to adopt policies that **"*comply with Union law on copyright and related rights*"**, which is necessary to make sure that no provider gains a competitive advantage **"*by applying lower copyright standards than those provided in the Union*"**. **Recital 108** [3] further mandates that GPAI providers should make **"*publicly available a summary of the content used for the training*"**, enabling the AI Office to verify that copyright laws, including opt-out rights under **"*Article 4(3) of Directive (EU) 2019/790*"** (**Recital 106** [3]), have been respected. Even open-source models are not exempt from these transparency duties if systemic risks exist.

Recital 105 [3] acknowledges the dual nature of generative AI. While it presents immense innovation potential, it also **"*challenges to artists, authors, and other creators and the way their creative content is created, distributed, used and consumed*"**, particularly when their content is scraped without permission. It reiterates that any use of copyrighted content **"*requires the authorisation of the rightsholder concerned unless relevant copyright exceptions and limitations apply*"** and that rights holders **"*may choose to reserve their rights over their works or other subject matter to prevent text and data mining, unless this is done for the purposes of scientific research*"**. However, where creators have expressly reserved their rights to opt out, **"*providers of general-purpose AI models need to obtain an authorisation from rightsholders if they want to carry out text and data mining over such works*"**.

To further support enforcement, **Recital 107** [3] calls for detailed and publicly accessible summaries **"*of the content used for training the general-purpose AI model*"** to facilitate rights enforcement by copyright holders, for example by **"*listing the main data collections or sets that went into training the model, such as large private or public databases or data archives, and by providing a narrative explanation about other data sources used*"**. At the same time, **Recital 109** [3] encourages simplified compliance pathways for SMEs and non-commercial users, taking into account **"*the size of the provider*"** and allowing **"*simplified ways of compliance for SMEs, including start-ups*"** as long as such measures do not come at the cost of creators' rights.

Ultimately, the tension between AI and art underscores a deeper question: not just whether AI *can* create, but what it means for humans if it *does*. Copyright protections are essential to defend artists' rights and creative integrity. However, in a digital world where content can be scraped, remixed, and repurposed in seconds, how far can those protections truly go? While large companies might have the legal resources to assert their rights and demand the removal of unauthorized content, what recourse does a small creator have when their work is quietly absorbed into a training dataset? If a piece of art is taken and repurposed without consent, has the artist not suffered economic, creative, or even emotional harm? As generative AI reshapes artistic production, frameworks like the AI Act must ensure that the burden of protection does not fall solely on those least equipped to defend themselves. In the end, it not only implies the future of art but also the dignity of those who make it.

5.4. The Singularity

The Singularity refers to a hypothesized moment when humans will merge with Artificial Intelligence, augmenting themselves with computational capacities far beyond the limitations of the biological brain. First introduced in a 1993 paper by computer scientist and science fiction writer Vernor Vinge, the concept described that "*the imminent creation by technology of entities with greater than human intelligence*" had led to "*the edge of change comparable to the rise of human life on Earth*" [25].

This vision has captivated inventors like Ray Kurzweil, who not only anticipates the arrival of the Singularity but actively plans to experience it (and, in this way, live forever). In his 2004 book *Fantastic Voyage: Live Long Enough to Live Forever* (Rodale, 2004), Kurzweil speculated that by the late 2020s, breakthroughs in AI and nanotechnology (which permits changes to the body at the cellular level) would allow humans to rebuild their bodies, overcome disease and deterioration, and eventually, death itself. He predicts that within the next decade, these technologies will extend lives at a faster rate than people age, reaching an "escape velocity" where life extension begins to outstrip the biological clock [23][24]. In this vision, death from ageing becomes optional, and surviving long enough to witness this moment could allow individuals to enter the Singularity and potentially live indefinitely.

Although this scenario remains speculative and may not yet warrant specific legislation, the convergence of AI with biological enhancement already raises ethical and regulatory questions. If human intelligence and life expectancy can be artificially expanded, who will have access to these capabilities? Can access to life itself (or even to immortality) be regulated? Should it be? And if so, can such access be democratized, or will it be reserved for the wealthy? Would the entities controlling these technologies be public institutions or private companies? Can something as fundamental as the right to live (or to enhance one's body) be governed by algorithms or be subjected to the decisions of private corporations?

While the AI Act does not directly address such possibilities (since, again, they are still just speculations), its principles could offer a potential starting point for future regulatory approaches. For example, its risk classification system (especially the provisions concerning high-risk systems in areas like health and safety) could one day be extended or adapted to cover neuro-enhancement tools or AI-powered biomedical implants, some of which are already beginning to emerge.

More broadly, the idea of the Singularity highlights a persistent challenge in the governance of emerging technologies, as innovation often moves faster than the development of legal and ethical norms. If frameworks like the AI Act are to remain effective in the long term, they must be not only robust but also flexible and anticipatory, capable of addressing present-day risks while preparing for deeper, more transformative changes. Addressing today's issues with foresight may help manage future scenarios before they stop being hypothetical.

As the line between biology and technology continues to blur, concepts like autonomy, consent, and access take on renewed significance. If it comes to the point where AI may allow extending a human's lifespan, who decides which lives are extended and on what basis? Should these life-altering technologies be treated as public goods, or will they be monopolized by private interests? In short, can immortality be regulated? And if so, by whom?

While technology (especially in medicine) has already extended human lives and saved many, a longer life is not necessarily a better one. These advancements already raise the question of limits, as eternal life (if ever achievable) risks becoming a nightmare in a world marked by deep inequality. Although the Singularity is not yet a reality (and may never be), it is not a future that can be entirely discarded, as it presents a scenario that forces us to confront the legal, moral, and social challenges that could arise if such a future becomes technically possible. Legal frameworks may not need to legislate for the Singularity today, but they should be prepared to engage with the possibilities it represents. In this sense, frameworks like the AI Act may serve not only to manage current risks, but also to lay the groundwork for governing the transformations of the future.

6. AI in politics and global security

6.1. The impact of AI regulation on democracy

The growing influence of AI systems in shaping public opinion and manipulating electoral outcomes has become a critical concern for democratic governance. During the already mentioned May 2023 U.S. Senate hearing, OpenAI's CEO, Sam Altman, voiced his unease about AI's potential to influence voters, comparing its possible influence to that of Google Search, which already shapes public perception by determining what content individuals see and how it is ranked [9]. Days later, while speaking in Europe, Altman acknowledged that AI-generated disinformation could play a role in the 2024 U.S. elections, although he suggested that social media platforms posed a greater threat in terms of distribution. "*You can generate all the disinformation you want with GPT-4, but if it's not being spread, it's not going to do much*" [11].

Altman's statements reflect broader concerns at the intersection of AI, media, and political influence, concerns that have already been addressed in the AI Act's **Recital 61** [3], which classifies **"*certain AI systems intended for the administration of justice and democratic processes*"** as high-risk due to their **"*potentially significant impact on democracy, the rule of law, individual freedoms*"**. Similarly, **Recital 110** [3] warns that GPAI models could pose systemic risks like **"*any actual or reasonably foreseeable negative effects on democratic processes, public and economic security*"** or **"*the dissemination of illegal, false, or discriminatory content*"**.

Historical precedents make these risks tangible. In March 2018, the Cambridge Analytica scandal exposed how massive datasets (illegally harvested from Facebook accounts) were used to build psychological profiles of voters. These profiles were then exploited to target individuals with personalized political ads, potentially influencing the outcomes of both the 2016 U.S. presidential election and the Brexit referendum. Initially employed by Texas Senator Ted Cruz and later by Donald Trump's election team, the data analytics company worked with researchers to develop "*a 120-question survey that seeks to probe personality*", asking all sorts of questions about personality and behaviour. The results, combined with polls, voter records, and online activity, were used in order to create personality models for targeting voters [72][73][74].

While some Facebook insiders claim that the impact of Cambridge Analytica was overstated, whistleblowers like Christopher Wylie insisted the company's misuse of Facebook data was decisive in Brexit's outcome and represented a form of "modern colonialism" in global elections. In his March 2028 testimony before the British Parliament, Wylie also contradicted his former boss, Alexander Nix, who stated to parliament members that the company had never used Facebook data. "*It is categorically false that Cambridge Analytica did not use Facebook data. It was their main database, and the algorithms were based on that*," Wylie told the U.K. Parliament's Digital, Culture, and Sport Committee. Regarding the June 2016 referendum on whether the country should remain in the European Union, which ended with a Brexit victory, Wylie was conclusive: "*I think it is completely reasonable to say there could have been a different outcome of the referendum had there not been, in my view, cheating*" [90].

In response to such practices, the AI Act introduces targeted regulatory mechanisms. **Annex III(8b)** [3] classifies as high-risk any **"*AI systems intended to be used for influencing the outcome of an election or referendum or the voting behaviour of natural persons in***

***the exercise of their vote in elections or referenda*"**, although this excludes uses related solely to **"*organise, optimise or structure political campaigns from an administrative or logistical point of view*"**. **Recital 136** [3] further clarifies that **"*the detection and disclosure that the outputs of those systems are artificially generated or manipulated*"** is essential, especially for **"*providers of very large online platforms or very large online search engines*"** in order to identify and mitigate risks **"*on democratic processes, civic discourse and electoral processes, including through disinformation*"**. These provisions aim to protect electoral integrity and civic participation.

Recent studies demonstrate just how urgent such protections have become. An investigation into Microsoft's Bing Chat related to the 2023 Swiss Federal Elections and 2023 State Elections in the German federal states of Hesse and Bavaria revealed that "*one third of Bing Chat's answers to election related questions contained factual errors*", ranging from incorrect election dates and outdated candidates to the fabrication of scandals. In many cases, these inaccuracies were wrongly attributed to credible sources, undermining public trust in media and institutional communication. Furthermore, even if the chatbot answered correctly to a prompt, it was not guaranteed that it would answer the same query correctly next time. Safeguards were also unevenly applied, which led to "*evasive answers 40% of the time*" and devalued the tool as a source of information. In multilingual contexts like Switzerland, Bing Chat performed even worse, demonstrating "*difficulties in dealing with different political systems*" and linguistic diversity **[80]**. Not only does this limit voters' access to accurate information, but it also biases AI-generated content toward more prominent political figures or narratives, disadvantaging lesser-known candidates and contributing to unequal political representation.

Transparency is another critical issue. AI systems like Bing Chat offer little clarity on how their responses are generated or what sources are used. As a result, seemingly credible but inaccurate answers may mislead users who mistake them for credible data without double-checking the sources. For voters, "*a search engine or chatbot powered by generative AI is not a reliable or transparent source of information*". For candidates, generative models "*can pose a risk to their reputation*". And for news outlets, "*made-up responses that cite their articles can undermine trust in their publications*" **[80]**. As generative AI becomes more widespread, its potential to disrupt reliable access to transparent and public information (a cornerstone of democracy) becomes more severe.

The philosophical implications are just as troubling. Belgian philosopher Mark Coeckelberg argues that democracy depends on informed citizens capable of rational debate and open to revising their views if necessary. Similarly to what was discussed in **Section 4.3** on misinformation, he warns that AI can bypass rational deliberation by targeting psychological vulnerabilities (automatically drafting opinion pieces or posting on social media), making it an especially effective tool of political propaganda. Unlike traditional media, AI-driven techniques are faster, more scalable, and capable of hyper-personalization. When individuals are manipulated without their awareness, both their individual autonomy and the foundational processes of democracy are compromised. As Coeckelberg asks, "*If your political beliefs are manipulated, what does your freedom to vote or your right to deliberate and participate in public discussions mean?*" **[2]**. AI literacy, as previously mentioned in **Recital 20** [3], becomes essential once again. Citizens must understand how these systems operate in order to responsibly engage with them.

These critiques echo historical warnings. Jean-Jacques Rousseau argued in *Discours sur l'origine et les fondements de l'inégalité parmi les hommes* (1755) that extreme economic inequality undermines democracy by letting the wealthy manipulate laws and public opinion.

More recently, Hannah Arendt warned that AI could alienate individuals and depersonalize political decision-making, making them not think about the consequences of their choices. She also cautioned that AI's ability to flood the public sphere with plausible falsehoods could foster an epistemic bubble where truth becomes indistinguishable from fiction, weakening democracy and contributing to creating the necessary conditions for the rise of totalitarianism [2].

Such instability erodes the trust necessary for collective democratic deliberation. When citizens are constantly exposed to conflicting information or manipulated through AI-driven profiling, when they are no longer able to tell apart facts from fiction, trust between citizens is destroyed, and democracy ceases to function. Even dataset creation is inherently political, as algorithmic profiling decisions about which data points are included or excluded are often made by opaque entities, distancing them from democratic control and turning them into technocratic choices rather than democratic ones [2].

Daniel Innerarity, professor of social and political philosophy, also reflects on the dual nature of algorithmic governance. He thinks it offers both objectivity (by redirecting "*certain debates to their more technical dimension*" and reducing "*the cult element of politics*" without eliminating the variety of opinions about that very same objectivity) and subjectivity (by "*identifying better our preferences and our interests*"), although this subjectivity is based not on us as individuals but on "*characteristics we share with population groups that we supposedly belong to*" [112]. Innerarity's analysis encapsulates the core democratic dilemma in AI governance. While algorithmic systems can offer apparent objectivity and personalization, they can also reinforce existing inequalities, induce artificial needs, and frame individuals through group-based profiling rather than as autonomous subjects.

In practice, this tension is already visible. In Austria, the public employment service (AMS) used an algorithmic profiling system to classify jobseekers into three categories (those with high chances to find a job within half a year, those with mediocre prospects on the job market, and those clients with a bad outlook of employment in the next 2 years) that determined the level of support they would receive, which raises ethical questions about discrimination and transparency [116]. In the U.S., the COMPAS algorithm evaluates the likelihood of a given defendant committing a crime after release. It uses a variety of factors (including the defendant's own responses to a lengthy questionnaire) to generate a recidivism-risk score between 1 and 10, which is then included in a defendant's PSI report supplied to the sentencing judge. As a result, a defendant's sentence is determined (to at least some degree) by COMPAS's recidivism risk assessment [117], which echoes dystopian films like *Minority Report* (2002). During the 2020 George Floyd protests, authorities in Minnesota were reportedly using contact tracing technologies, facial recognition software, and cellphone tracking to monitor demonstrators, exemplifying how surveillance tools intended for public health can be repurposed for political control [118]. In China, South Korea, Singapore, and Taiwan, AI was used to monitor infected individuals during the COVID-19 pandemic via mobile phones, raising concerns about mass surveillance and the erosion of civil liberties [119].

During the February 2025 *HispanIA 2040* conference, Spanish Prime Minister Pedro Sánchez warned against allowing Silicon Valley tech corporations to shape the future of public debate and government action. He noted that throughout history, these technologies tend to reinforce the *status quo*, making the rich richer and reinforcing inequality. At the same event, Sánchez also announced the creation of *Alia*, an ensemble of public, open-source AI models developed in Spanish as well as in Spain's co-official languages [28]. This initiative aims to create an inclusive and culturally representative technological

infrastructure that is aligned with public values, especially in a world dominated by English-language models. As the Bing Chat performance in multilingual contexts like Switzerland showed, the lack of linguistic diversity in AI systems can lead to the underrepresentation of entire communities, distorting political discourse and decision-making.

Such efforts reflect a broader global need. Coeckelberg stresses that isolated national policies are insufficient when data and algorithms easily transcend territorial boundaries, and that efforts to regulate these systems must be harmonized across borders. However, global coordination also raises concerns about technocratic overreach and the risk of authoritarianism **[2]**. Organizations like the Norwegian Open AI Lab or private companies like Glovo emphasize the need to define "*a set of common regulatory principles across applications of technology for autonomous systems*" and to "*align AI usage standards with EU principles*". The Act's ambition to position the EU as a leader in AI governance is further reflected in global dialogues, such as the EU-U.S. Trade and Technology Council **[62]**, which aims to harmonize transatlantic AI policy and promote international cooperation.

AI regulation sits at the core of a profound democratic dilemma. While these technologies can enhance administrative efficiency and even support civic participation, they also pose serious risks to individual autonomy, public trust, and political equality (especially when used to manipulate public opinion or spread misinformation, a matter already discussed in **Section 4.3**). One of the most significant threats AI poses to democracy is its capacity to distort the information ecosystem, erode rational debate, and influence electoral outcomes. The AI Act acknowledges these dangers by classifying systems intended to influence voting behaviour or public opinion as high-risk. However, in a world where the Cambridge Analytica scandal has already shown the real-world consequences of data-driven manipulation, legal classification alone is not enough. To truly protect democratic integrity, clear consequences and enforceable sanctions for misuse must accompany regulation. Truthful, accessible information is foundational to democratic life, and safeguarding it must be both a political and legal priority.

Transparency in AI-generated content and strong AI literacy are essential in this regard. While the AI Act rightly emphasizes the need for public understanding, this must translate into widespread education that empowers citizens to question, verify, and critically engage with AI systems. If disinformation becomes so pervasive that truth is indistinguishable from falsehood, the door opens to authoritarianism, and the best defence against this is a well-informed public that understands the technologies shaping its reality. Moreover, since AI must be designed to serve diverse populations, democracy should not be limited to a single linguistic or cultural framework. Supporting the development of open, public AI models in multiple languages (like Spain's *Alia* initiative) is essential for ensuring inclusive participation and fair representation in the digital age.

Control over AI infrastructures also raises fundamental democratic concerns. If these tools remain in the hands of a few powerful private tech companies, public discourse risks becoming a tool of economic elites. As Rousseau and Arendt warned in their respective times, concentrated power (whether economic or technological) can easily be turned into a tool for manipulation and domination. Safeguards like the AI Act's ban on biometric surveillance, including facial recognition, are critical for protecting civil liberties.

Still, no matter how robust the AI Act may be within the EU, its effectiveness will also depend on proper enforcement and international cooperation, with initiatives like the EU-U.S. Trade and Technology Council representing a necessary step towards aligning AI governance frameworks globally. Without shared standards and cross-border cooperation, national or

regional laws may fall short, allowing technological abuses to slip through jurisdictional gaps, weakening democratic safeguards.
In the end, ensuring that AI supports rather than undermines democracy requires not just legal frameworks like the AI Act, but a broader cultural transformation, one that embeds reflection into the design process, cultivates AI literacy among citizens, and demands democratic control over technological infrastructures. As Coeckelberg argues, AI alone is unlikely to destroy democracy by itself unless democracy is already in trouble, but it can accelerate its decline if democratic institutions are already weak [2]. Strengthening those institutions and making sure that the public is equipped to engage with them is the best way to make sure that AI becomes a tool that supports democracy rather than erodes it.

6.2. The intersection of AI, war, and global security

As AI becomes more powerful and accessible, its potential misuse in military and security contexts has emerged as a major global concern. In February 2025, former Google CEO Eric Schmidt warned that countries like "*North Korea, or Iran, or even Russia*" could exploit AI to develop weapons of mass destruction, including biological weapons. He emphasized the need for governmental oversight of private tech companies developing AI models but also cautioned against overregulation (particularly in Europe), arguing that it could suppress innovation. In his own words, "*the AI revolution, which is the most important revolution in my opinion since electricity, is not going to be invented in Europe*" if regulation goes too far [46]. This highlights a critical tension that the AI Act must carefully navigate.
That same month, Alphabet (Google's parent company) revised its AI policy, removing explicit commitments to rule out applications "*likely to cause harm*". This opened the door to allow the use of AI in the development of weapons and surveillance tools. Human Rights Watch criticized this shift, arguing that AI complicates accountability in battlefield scenarios that "*may have life or death consequences*", and stressed that voluntary guidelines cannot replace binding laws. The ongoing integration of AI in military operations (such as in Ukraine and the Middle East, where it is being integrated into targeting systems and battlefield logistics) has amplified calls for stricter regulations. Campaigners have expressed alarm about AI potentially making life-or-death decisions autonomously, underscoring the urgent need for clear and enforceable safeguards [59].
International security experts like Alexander Kmentt, Director of the Disarmament, Arms Control and Non-Proliferation Department of the Austrian Foreign Ministry, have expressed alarm over the rapid development of autonomous weapons systems (AWS), warning that "*humanity is about to cross a threshold of absolutely critical importance (...). This window [to regulate] is closing fast*". Some AI-enabled weapons are already being sold without autonomy limits, leaving clients to decide whether machines can fire independently, without human input. For example, an Israeli weapons system reportedly identified people as threats based on the presence of a firearm (though, like humans, these systems are not immune to errors). While companies like Athena AI claim their systems are "*designed for AI on the loop [with a human operator always involved]*" (which would mean that "*AI does not make targeting decisions*"), critics like Catherine Connolly, who holds a PhD in international law and security studies, warn that switching to full autonomy might require nothing more than a minor software update. This threat has mobilized campaigns like *Stop Killer Robots*, which advocate for an international treaty to ensure "*meaningful human control over systems that detect and apply force to a target based on sensor inputs rather than an immediate human command*", including applications beyond warfare, such as policing and border control [60].

Despite growing concerns, the AI Act does not directly regulate AI used for military, defence, or national security purposes. **Recital 24** [3] clearly states that such uses fall outside the scope of the regulation **"*regardless of which type of entity is carrying out those activities, such as whether it is a public or private entity*"**. The justification for excluding military and defence applications lies in both **"*Article 4(2) TEU*"** and **"*the specificities of the Member States' and the common Union defence policy covered by Chapter 2 of Title V TEU that are subject to public international law*"**, while the exclusion of national security applications is grounded in **"*the fact that national security remains the sole responsibility of Member States in accordance with Article 4(2) TEU*"**. However, when an AI system initially designed for military or security purposes is used **"*temporarily or permanently for other purposes, for example, civilian or humanitarian purposes, law enforcement or public security purposes, such a system would fall within the scope of this Regulation*"** and compliance with the AI Act would be mandatory. This distinction highlights the AI Act's horizontal nature as a general product regulation, while war-related AI remains subject to specialized legal frameworks under international law.

Still, this separation between civilian and military AI governance has limitations. Past controversies, like Google's 2018 involvement in Project Maven (a Pentagon initiative to use AI for processing video footage), show how blurred the boundaries between supportive and offensive military applications can be. Although Google assured that its contribution was non-offensive, nearly 3,100 employees published an open letter to CEO Sundar Pichai demanding an end to the collaboration. The AI system was capable of automatically detecting and classifying elements within images, potentially enhancing drone targeting capabilities [89]. The Pentagon stated that "*we are in an AI arms race*" [105], and although the project is initially intended to assist human analysts, its integration into warfare workflows raises fears that AI may incrementally assume more lethal decision-making roles [89].

These fears are not theoretical. In August 2017, 110 leaders from AI and robotics companies worldwide signed an open letter to the UN's Group of Governmental Experts, urging international regulation of lethal autonomous weapons systems (LAWS). They warned that such technologies could constitute "*the third revolution in warfare*", rapidly increasing the scope and speed of armed conflicts if developed and subsequently deployed without restrictions. Such concerns are far from hypothetical. Autonomous weapon systems were already being deployed as early as the mid-20th century (originally for surveillance purposes), and have since evolved into combat-ready systems used in real-world conflicts. For instance, the Israeli army deployed such weaponry in Lebanon, and in the early 2000s, the U.S. deployed Predator drones equipped with missiles in Afghanistan and Libya. More recently, the UN Panel of Experts on Libya reported that a Turkish STM Kargu-2 drone was used against Hafter Affiliated Forces, noting that it was "*programmed to attack targets without requiring data connectivity between the operator and the munition*" [108].

In response, UN Secretary-General António Guterres has called for a legally binding international agreement that would "*prohibit lethal autonomous weapon systems that function without human control or oversight, and which cannot be used in compliance with international humanitarian law*". He emphasized the humanitarian, legal, ethical, and security problems, warning that such systems threaten basic human rights and freedoms [108].

According to the European Parliament's 2018 Resolution 2018/2752(RSP) [109], LAWS are defined as "*weapon systems without meaningful human control over the critical functions of selecting and attacking individual targets*", although excluding "*non-autonomous systems such as automated, remotely operated and tele-operated systems*" from this category, such as remotely operated drones. Despite the rapid evolution of AI and robotics, this definition

has not been updated since 2018, and the international community has yet to reach a consensus on a unanimous definition of the LAWS. Scholars like Mariarosaria Taddeo and Alexander Blanchard advocate for a broader, value-neutral definition that incorporates characteristics such as "*autonomy, adaptability of AWS, human control, and purpose of use*" [108]. They argue that the EU's outdated approach weakens the Union's regulatory effectiveness. Nonetheless, it is also true that the EU legislation opposes to the funding, production, and procurement of LAWS, as the legislation explicitly prohibits funding for "*actions related to the production of lethal autonomous weapons without the possibility of meaningful human control over selection and engagement decisions when carrying out strikes against humans*" [110], aligning with the Secretary-General's call for ethical limits on AI-enabled warfare.

The ethical dilemma at the heart of LAWS is the delegation of life-or-death decisions to machines. The UN maintains that delegating *the decision to take a human life by machines*" is inherently dehumanizing, as such systems lack "*empathy, compassion and the ability for moral reasoning*" and could give rise to "*unjustified violence and civilian casualties*". In the ongoing genocide in Gaza, Israeli systems like "The Gospel" and "Lavender" have reportedly been used by the Israeli Defence Forces to identify Hamas members, causing widespread destruction and substantial civilian casualties. Although these systems do not yet qualify as fully autonomous, their reliance on minimal human oversight exemplifies the inadequacy of current safeguards [108].

Moreover, LAWS challenge compliance with International Humanitarian Law (IHL), especially the principle of distinction, which requires that combatants must be clearly distinguished from civilians, and that attacks must only target legitimate military objectives. In urban or asymmetric warfare, this becomes increasingly difficult to uphold when targeting decisions are left to algorithms. While the AWS Resolution acknowledges accountability as an important issue, it fails to provide specific regulations on the issue. Scholars have proposed concrete steps, such as requiring all LAWS implemented in EU member states to include accountability mechanisms, like, for example, target selection data logging, the justification for particular actions, and system-human operator communication logs. Without clear attribution of responsibility, violations of IHL may go unpunished [108].

The AWS Resolution emphasizes that "*human involvement and oversight are central to the lethal decision-making process, since it is humans who remain accountable for decisions concerning life and death*" [109], yet the term "meaningful human control" remains vague. Traditionally, individuals such as combatants, commanders, or political leaders have been held accountable for violations of IHL, but when autonomous systems operate and act independently, it becomes unclear who is liable, especially if the use of force was not explicitly authorized by a human operator. The lack of a precise definition creates legal loopholes that enable the deployment of potentially harmful technologies without adequate oversight, ultimately eroding accountability. As a result, "meaningful human control" risks becoming a rhetorical shield rather than a functional barrier to autonomous lethality, undermining legal accountability and potentially enabling violations of IHL and IHRL (International Human Rights Law) to go unpunished. As such, the EU must urgently review and update its 2018 definition of LAWS. A modernized framework should incorporate autonomy and adaptability while clearly distinguishing between AWS and LAWS, reflecting their differing ethical and operational implications [108]. Doing so would improve regulatory clarity and reinforce the EU's commitment to international law and human dignity.

From a humanitarian perspective, the deployment of LAWS should never result in indiscriminate harm or target combatants *hors de combat* (incapable of performing their

combat duties during war). While the AWS Resolution distinguishes between AWS and defensive systems excluded from the definition of AWS, it fails to address the offensive use of LAWS. The EU should clarify the conditions under which LAWS may be used in armed conflict, with the main criterion being that LAWS could only be used in areas with no civilian presence, limiting its use to military objectives. EU legislation must also ensure that any deployment complies with IHL, particularly in relation to the protection of non-combatants, medical staff, and humanitarian workers **[108]**.

The rapidly evolving nature of AI in warfare demands urgent and coordinated international action. While the AI Act establishes a framework for civilian oversight, it leaves military applications unregulated, a separation that is increasingly unsustainable. The deployment of AI in armed conflicts is far too important and urgent to be governed by vague concepts or outdated definitions. Lives are at stake.

The issue of liability (already discussed in **Section 4.1**) resurfaces with critical importance in the context of AI-enabled weapons. Without clear attribution of responsibility, violations of IHL may go unpunished. Similarly, the concept of “meaningful human control”, though central to the current legal framework, also remains dangerously ambiguous. In the high-stakes environment of warfare, such vagueness can easily become a legal loophole, effectively erasing accountability for the loss of human life. To address this, human oversight and control must be guaranteed through robust accountability mechanisms, such as data logging of target selection, justification of actions, and system-human interaction, as proposed by experts. These tools are essential to trace actions back to individuals or institutions and ensure compliance with IHL.

Each day without updated or specific regulation leads to preventable deaths in armed conflicts due to AWS operating without human judgment or legal oversight, deaths for which no one may ultimately be held accountable due to legal voids that must not be allowed to persist. The EU must modernize its legal instruments to ensure that accountability, human rights, and international law are not casualties of algorithmic warfare, determining not only the EU’s leadership in ethical AI but also the broader contours of 21st-century warfare.

7. Future implications and open questions

7.1. Adaptation and flexibility of AI regulation

During the previously mentioned U.S. Senate hearing on May 16th, 2023, concerns were raised about the staggering pace at which AI is developing. Republican Senator Josh Hawley referenced a widely publicized open letter (signed by prominent figures in the AI field) calling for a pause in AI development. He used this to propose a general six-month pause, a suggestion that was met with hesitation from the hearing's witnesses. In response, Senator Blumenthal, the chair of the hearing, warned against complacency, stressing that "*the world won't wait*" and that "*sticking our head in the sand*" is not the solution [9]. Just days later, a group of 19 current and former leaders of the Association for the Advancement of Artificial Intelligence (a 40-year-old academic society) released their own letter, warning about the growing risks posed by AI [12].

Still, some voices in the field argue that even a temporary halt would fall short. Eliezer Yudkowsky, decision theorist and researcher at the Machine Intelligence Research Institute (MIRI), argues that the real danger lies not just in achieving human-level AI, but in what comes afterward. He warns that once we cross the threshold into superhuman intelligence, the associated risks become existential. According to Yudkowsky, we currently lack the scientific understanding and precision required to safely align such systems with human values. "*Without that precision and preparation,*" he cautions, "*the most likely outcome is AI that does not do what we want, and does not care for us nor for sentient life in general*", a scenario that could lead to catastrophic consequences. For this reason, he calls for a global, indefinite moratorium without exceptions, not even for governments or militaries [35].

These growing concerns underscore the urgent need for clear, enforceable, and forward-looking legislation, which is something the AI Act aims to provide. However, it feels unavoidable to ask, can legal frameworks evolve quickly enough to govern such a rapidly evolving field?

To address this, some experts advocate for regulatory sandboxes, controlled environments where new technologies can be tested under the guidance of regulators. By offering space for experimentation, these sandboxes aim to bridge the gap between broad legal principles and the technical complexities of AI compliance. This approach could also enable governments to learn and adapt alongside the private sector, potentially leading to the development of cohesive international standards and a global AI policy direction [62].

In parallel, the AI Act includes formal mechanisms for periodic review and adjustment. **Recital 174** [3] acknowledges the rapid pace at which technologies advance and sets a structured evaluation schedule, specifying that **"*the Commission should evaluate and review this Regulation by 2 August 2029 and every four years thereafter and report to the European Parliament and the Council*"**. Additionally, it mandates that **"*the Commission should carry out an assessment of the need to amend the list of high-risk AI systems and the list of prohibited practices once a year*"**. Starting on **"*2 August 2028 and every four years thereafter*"**, the Commission is also required to evaluate **"*the need to amend the list of high-risk areas headings in the annex to this Regulation, the AI systems within the scope of the transparency obligations, the effectiveness of the supervision and governance system and the progress on the development of standardisation deliverables on energy efficient development of general-purpose AI models, including the need for further measures or actions*"**.

Furthermore, **"*by 2 August 2028 and every three years thereafter*"**, the Commission should evaluate **"*the impact and effectiveness of voluntary codes of conduct*"**.
This commitment to continuous oversight is reinforced by **Article 112(3)** [3], which states that by **"*2 August 2029 and every four years thereafter, the Commission shall submit a report on the evaluation and review of this Regulation to the European Parliament and to the Council*"**. This report must include **"*an assessment with regard to the structure of enforcement and the possible need for a Union agency to resolve any identified shortcomings*"**. When necessary, it may also include proposals to amend the Regulation, and it must be made public. These scheduled reviews are intended to ensure that regulation remains both relevant and responsive to ongoing technological developments.
Recital 65 [3] further emphasizes that the risk-based regulatory framework that characterizes the AI Act (especially for high-risk AI systems) must be **"*regularly reviewed and updated to ensure its continuing effectiveness*"**, that proper justification and documentation must be included for any significant decisions taken under the Regulation. Similarly, **Article 96(2)** [3] provides that **"*the Commission shall update guidelines previously adopted when deemed necessary*"** at the request of **"*the Member States or the AI Office, or on its own initiative*"**.
Altogether, these provisions demonstrate that the AI Act is aware of the rapid pace of technological progress and recognizes the corresponding need for regulatory adaptability. By embedding mandatory review and update mechanisms within the legislation itself, the Act reflects a clear predisposition toward adaptability. However, whether the selected revision intervals (every three to four years) are truly sufficient remains uncertain. Given the current speed of AI development, these periods may be proven to be too slow. Although the existence of periodic reviews is encouraging, it remains to be seen whether they will allow for timely and adequate responses, a question that may only be answered after the first review cycle is completed. What happens if AI advances significantly between one revision and the next?
As technological acceleration continues, legal frameworks must not only keep pace but also be structurally equipped with the necessary flexibility to respond to change in near real-time. Although the AI Act is designed for periodic revision, the very need for scheduled updates highlights the challenges of creating a regulation that is both robust and adaptable from the beginning. As the pace of technological change accelerates, legislative responses must do more than just react, as they must also anticipate emerging risks and remain structurally capable of rapid, informed adaptation. The extent to which the AI Act can accomplish this will ultimately determine its effectiveness as a global model for trustworthy AI governance.

7.2. Digital abuse in times of AI

Article 3 [3] of the AI Act defines **"*deep fake*"** as **"*AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places, entities or events and would falsely appear to a person to be authentic or truthful*"**. A clear example of the risks associated with this technology was presented during the 2023 U.S. Senate hearing on AI regulation, where Connecticut Senator Richard Blumenthal opened with an audio clip that mimicked his own voice. The clip had been entirely generated by AI, using ChatGPT to draft remarks based on his previous speeches. While the demonstration was technically impressive, it served as a clear illustration of the growing dangers tied to generative AI, such as voice cloning, misinformation, and deepfakes [9].

These risks are no longer hypothetical. Across the globe, individuals (especially women) are suffering from the malicious use of AI-generated deepfake content. One harrowing case involved an Australian woman named Hannah, who discovered that her social media photos had been stolen and used to create hundreds of sexually explicit deepfake images, which were later shared online in a public forum filled with threats and harassment. At the time (2022), legal loopholes in New South Wales meant that no criminal charges were applicable for such offences, and Australia's eSafety Commissioner could only assist in removing the content. Although Andy, Hannah's long-time friend and the creator of the deepfakes, was arrested, there were no laws in place to prosecute him for creating the deepfakes themselves. Australia has since updated its legislation, criminalizing deepfake pornography nationwide in 2024, but the case highlighted the gap between technological advances and legal protections. As Hannah herself put it, "*for me, and for the other girls, it is forever… they will always be on the internet (...). He had turned every single one of those moments for us into porn. And so when you see that photo… well, now I see myself getting raped*" **[54]**.

The U.K. faces a similar legal gap. In 2023, the Online Safety Bill made it illegal to share or threaten to share sexually explicit deepfake images in England and Wales. However, the bill was drafted at a time when deepfake creation required technical expertise. Today, it only takes access to an app. This shift in accessibility means much of the legislation is already outdated, often forcing victims to act as investigators themselves, collecting evidence, tracking down content, and pushing platforms to take it down. While sharing deepfake pornography is now illegal in parts of the United Kingdom, creating or requesting such content remains unpunished under current law. In 2024, conservative peer Baroness Owen proposed a law to criminalize the creation or solicitation of intimate images without consent, highlighting that 99% of explicit deepfakes target women. Her proposal would establish new offences with penalties of fines and up to six months in jail. However, the government declined to support the bill, prompting Baroness Owen to criticize ministers for "*delaying action*", stating that this was "*a betrayal of those who need our protection the most*". Alongside other campaigners like Baroness Kidron, she has called for consent-based legislation that does not require victims to prove malicious intent, an evidentiary burden that makes prosecution incredibly difficult **[55][56]**.

The demand for deepfake content is alarming and deeply troubling. According to the cybersecurity company Deeptrace, 96% of all deepfakes are non-consensual pornography, including revenge porn. In online forums, users submit requests for deepfake videos of their wives, neighbours, co-workers, and even their mothers, daughters, and cousins. Content creators respond with step-by-step instructions on what source material is needed, advice on which filming angles work best, and how much it will cost **[56]**.

South Korea faces an even more severe crisis. The country struggles with an epidemic of digital sex crimes, with hundreds of women and girls targeted through deepfake sexual images being shared online. Between 2021 and September 2024, reported cases of deepfake-related offences soared from 156 to 812, with 387 suspects apprehended (83.7% of whom were minors). Thousands of users participate in online forums dedicated to producing and sharing AI-generated pornography, often targeting women and girls as young as high school and middle school students. Investigations have uncovered entire online groups focused on specific victims, with users sharing photos on Telegram of women they knew and using AI software to convert them into fake pornographic images within seconds. The public indignation that followed led to legislative reform, with the Korean government establishing a centre to assist survivors of digital sex crimes. In September 2024, Korea's National Assembly and Cabinet amended the *Act on Special Cases Concerning the*

Punishment of Sexual Crimes, a bill that included mandatory minimum sentences (including one year for blackmail with deepfake material and three years for producing deepfake material for dissemination), raising the maximum sentence for producing and distributing deepfake pornographic material, and criminalizing the creation, possession and viewing of deepfake material [86][87][88].

In contrast, the AI Act's approach on this matter relies more on transparency obligations than on liability for the creation of such content. As previously noted in **Section 4.3**, **Article 50(4)** [3] requires deployers of an AI system **"*that generates or manipulates image, audio or video content constituting a deep fake*"** to disclose that **"*the content has been artificially generated or manipulated*"**, although exceptions apply in certain cases like when **"*the use is authorised by law to detect, prevent, investigate or prosecute criminal offence*"**. Where **"*the content forms part of an evidently artistic, creative, satirical, fictional or analogous work or programme*"**, disclosure must still occur, but **"*in an appropriate manner that does not hamper the display or enjoyment of the work*"**.

As already mentioned in **Section 4.3**, **Recital 133** [3] underscores the gravity of the situation, stressing that AI systems can now generate **"*large quantities of synthetic content that becomes increasingly hard for humans to distinguish from human-generated and authentic content*"**. This poses serious threats to **"*the integrity and trust in the information ecosystem, raising new risks of misinformation and manipulation at scale, fraud, impersonation and consumer deception*"**. To mitigate this, the AI Act requires providers of those systems to **"*embed technical solutions that enable marking in a machine readable format and detection that the output has been generated or manipulated by an AI system and not a human*"** such as, for example, **"*watermarks, metadata identifications, cryptographic methods for proving provenance and authenticity of content, logging methods, fingerprints or other techniques, as may be appropriate*"**. **Recital 120** [3] ties these transparency obligations to the EU's broader digital governance efforts. These obligations **"*are particularly relevant to facilitate the effective implementation of Regulation (EU) 2022/2065*"**, especially for **"*providers of very large online platforms or very large online search engines*"** that must **"*identify and mitigate systemic risks that may arise from the dissemination of content that has been artificially generated or manipulated, in particular risk of the actual or foreseeable negative effects on democratic processes, civic discourse and electoral processes, including through disinformation*"**.

While the AI Act provides an important foundational framework, its scope remains limited to transparency requirements. It ensures that users know when content has been AI-generated or manipulated, but what happens to the victims? What legal recourse exists if AI-generated images of them are distributed without consent? Who protects them, and what penalties do perpetrators face?

As the cases from Australia, South Korea, and the U.K. clearly show, national laws are still struggling to catch up. The AI Act does not directly address the criminal or civil liability for deepfake abuse, nor does it offer support mechanisms for victims. Laws may take years to draft, but deepfakes take mere seconds to create, and their consequences can last a lifetime. This underscores the urgent need for complementary legal measures (if not within the Act itself, then through parallel legislation) that establish proactive, consent-based protections, define clear penalties, and ensure comprehensive support for those affected. Without such measures, the dangers of deepfake abuse will continue to escalate, threatening the very dignity, autonomy, and safety that digital regulation seeks to uphold. The longer this legislative gap remains, the more lives it will irreparably damage.

8. Conclusions

The European Union's AI Act represents one of the most ambitious efforts to date to regulate AI in a way that balances innovation with the protection of fundamental rights. As the analysis conducted throughout this final project has demonstrated, the legislation addresses the increasingly urgent need to ensure that AI systems (already deeply embedded in sectors such as education, healthcare, labour, and politics) are developed and deployed ethically, transparently, and inclusively. The Act introduces a risk-based framework that classifies AI systems into four categories (minimal, limited, high, and unacceptable risk) and regulates them accordingly. However, the rapid evolution of AI technologies raises a crucial question: is this regulation enough? And how often should it be revised and updated to remain effective and relevant?

A key insight from this research is that while the AI Act establishes a strong foundation (particularly through provisions on workers' rights, AI literacy, and data quality), it leaves several critical gaps unaddressed. These include unresolved issues around legal liability, the absence of binding environmental obligations, and the systemic invisibility of the hidden labour that supports AI development. Although the Act introduces a set of regulatory controls (both pre-deployment and post-deployment) primarily aimed at AI-based systems and their industry, it lacks sufficient mechanisms for addressing harms that take place after deployment, leaving affected individuals without clear paths to redress in cases of discrimination, misinformation, or emotional harm.

As IBM recognized decades ago [98], systems that cannot bear responsibility should not be granted the authority to make consequential decisions. Today, we face a modern-day trolley problem. On one track lie AI's promises, including its potential for innovation, offering emotional support, and improving the quality of life. On the other, the dangers of malfunctioning systems, unregulated platforms, and vulnerable users whose lives may be impacted by artificial agents that are not (and cannot be) held accountable. In this analogy, AI is the unseen hand at the lever, programmed by developers who remain largely shielded from responsibility. Systems that cannot be trusted to function reliably should not be allowed to impact or influence human lives in the first place, and in critical areas such as immigration, education, and public safety, the absence of adequate regulation allows too much room for harm and too little opportunity for redress. The question is no longer just *who* pulls the lever, but whether society truly *understands* the consequences of the tracks.

In education, AI presents both promise and peril. While it can enhance personalized learning and improve accessibility, its uncritical use risks undermining academic integrity, eroding critical thinking, and widening educational inequality (especially among students who cannot access or afford such technologies). Although the AI Act acknowledges these risks and opportunities, its effectiveness will ultimately depend on how national governments and educational institutions translate its principles into practical, inclusive, and actionable strategies. Pedagogy, institutional policy, and AI literacy must play equal roles and work in tandem.

The intersection of AI and democracy also demands urgent attention. Generative models can mislead, manipulate, and polarize, as evidenced by the proliferation of deepfakes and algorithmically amplified echo chambers. The danger lies not only in the realism of AI-generated content but also in the public's growing inability to distinguish truth from fabrication. While the AI Act introduces transparency obligations, its enforcement mechanisms remain vague. A stronger accountability model is needed, one that involves

developers, platforms, and a society equipped to recognize and resist manipulation. As shown in **Section 6.1** through the analysis of Microsoft's Bing Chat [80], the public release of generative AI tools without adequate safeguards threatens a cornerstone of healthy democracies: access to reliable information. If such systems continue to be deployed, companies must implement stronger safety measures, and public regulators must enforce binding regulations for their development, deployment, and accountability of generative AI.

Transparency must also be user-centred. While the Act mandates explainability, it remains unclear whether most users can meaningfully understand how AI systems affect their lives and data. Legislation must go beyond technical transparency and ensure *comprehensible* transparency, allowing non-experts to truly understand and challenge decisions, especially when explanations are overly technical or opaque.

AI is not neutral, as the datasets used to train it reflect the power dynamics and priorities of those who design and deploy it. As AI investigator Yi Zeng observed, the issue lies not in AI itself, but in its involvement with ideological and social structures, with the key being that "*AI researchers should have an ethical awareness when designing and developing the technology*" [121]. After all, there is no such thing as "AI in itself", since it is always connected to broader human systems [2]. Without strong safeguards, AI could become a tool for surveillance, censorship, and manipulation, concentrating power in the hands of a few. Only transparent governance and civic participation can ensure that AI strengthens, rather than weakens, democratic institutions.

Environmental sustainability is also insufficiently addressed. Although ecological concerns are acknowledged, the Act fails to impose concrete and enforceable sustainability standards. As AI's energy consumption and carbon emissions continue to rise (particularly through data centres and training infrastructure), the absence of binding environmental provisions risks reducing the Act's ecological ambitions to mere symbolism. At the same time, the labour landscape is transforming rapidly. While AI holds promise for the creation of jobs that cannot even be imagined yet, automation also threatens to displace millions of workers while concentrating wealth and control in fewer hands. The Act addresses some of these issues through provisions on workers' rights, AI literacy, and training in the use of these new technologies. However, it fails to protect workers who make AI possible, particularly workers from the Global South tasked with filtering harmful content used for the creation of datasets, often at significant emotional cost.

Even more alarming is the development of autonomous weapons, systems capable of making kill decisions without human oversight. As stated in **Section 6.2**, lives are at stake, and delegating such life-or-death decisions to algorithms dehumanizes violence and erodes accountability. While AWS and LAWS fall outside the AI Act's scope, the need for human oversight and international legal standards in this area cannot be ignored.

These concerns extend to AI systems capable of self-improvement. As models evolve beyond their original design, new regulatory dilemmas arise. How can systems that mutate in complexity, or exhibit behaviour their developers did not anticipate, be governed? How can accountability be maintained when systems evolve autonomously? Legal frameworks must be flexible and anticipatory, designed not only for current technologies but also for those yet to emerge. The Singularity theory, discussed in **Section 5.4**, serves as a reminder of technology's endless and unpredictable possibilities. Even if such futures seem distant and almost impossible, laws must be adaptable enough to respond to outcomes that cannot even be imagined yet.

Lastly, can a machine make moral decisions? Who programs its values? What happens when algorithms must choose between two harmful outcomes? Ethical responsibility cannot

be delegated to systems that lack consciousness, empathy, or moral reasoning. AI must be developed in accordance with publicly debated ethical principles and always under democratic oversight. As Alan Turing noted in 1950, "*we can only see a short distance ahead, but we can see plenty there that needs to be done*" **[120]**. That sense of urgency still resonates. However, as Coeckelberg reminds, technology does not *necessarily* lead to centralized power, oppression, or violence, which means there is still hope for emancipation and social change **[2]**. The real question is then whether technology will continue to be used for repression and exploitation, or whether something different can be imagined and built.

To summarize, the Act falls short in several key areas. It emphasizes environmental protection but lacks binding mechanisms to address environmental sustainability. It fails on liability matters by not providing robust *ex post* remedies for those harmed by AI systems. In education, its current formulation promotes accessibility and AI literacy but does not fully address ethical and pedagogical dilemmas. It mandates technical transparency but does not guarantee comprehensible transparency for non-expert users, who should not be expected to understand too technical explanations for their knowledge. It recognizes displaced labour yet remains silent on the exploitation of dataset content moderators. It requires deepfake identification but lacks legal remedies for victims of such abuse, which disproportionately targets women. And even if it explicitly excludes autonomous weapons from its scope, the EU cannot ignore the urgent need for updated and dedicated legislation in this area.

In light of all this, the sufficiency of the AI Act must be approached with cautious optimism and critical realism. Its existence marks a significant political achievement, especially in a global context where approaches to AI regulation vary widely across countries like the U.S., Canada, or the U.K. However, that does not make it sufficient. The AI Act must not be understood as a final solution, but as a vital first step: really ambitious in scope, but inevitably limited as a first attempt. Flexibility is built into the Act through its review mechanisms (with the first revision scheduled for 2028), but whether that interval is adequate or too long remains to be seen. Ultimately, the AI Act's long-term relevance and success will depend on the EU's willingness to revise it regularly, enforce it rigorously, and adapt it responsively, always with active involvement from civil society. Because in the age of algorithms, the future will not be written by machines, but by the clarity of our laws, the strength of our institutions, and our collective commitment to ensure technology serves the public good.

References

- **[1] BAUMAN, Z. (2000).** *Modernidad líquida*. Cambridge: Polity Press and Blackwell Publishers Ltd., p. 198.
- **[2] COECKELBERGH, M (2024).** *Por qué la IA debilita la democracia y qué hacer al respecto*. Cambridge: Polity Press Ltd.
- **[3] EUROPEAN UNION.** REGULATION (EU) 2024/1689 OF THE EUROPEAN PARLIAMENT AND OF THE COUNCIL of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act). *Official Journal of the European Union*, 12 July 2024, No. 1689. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689> [Consultation: 7 April 2025]
- **[4] EUROPEAN PARLIAMENT.** *EU AI Act: first regulation on artificial intelligence*. <https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence> [Consultation: 17 April 2025]
- **[5] EUROPEAN COUNCIL.** *Artificial intelligence (AI) act: Council gives final green light to the first worldwide rules on AI*. <https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/> [Consultation: 17 April 2025]
- **[6] EU ARTIFICIAL INTELLIGENCE ACT.** *The AI Act Explorer.* <https://artificialintelligenceact.eu/ai-act-explorer/> [Consultation: 7 April 2025]
- **[7] EU ARTIFICIAL INTELLIGENCE ACT.** *High-level summary of the AI Act.* <https://artificialintelligenceact.eu/high-level-summary/> [Consultation: 7 April 2025]
- **[8] TECHCRUNCH.** *Sam Altman's big European tour.* <https://techcrunch.com/2023/05/25/sam-altman-european-tour/> [Consultation: 13 February 2025]
- **[9] BBC.** *ChatGPT boss says government regulation of AI is 'crucial'*. <https://www.bbc.com/news/live/world-us-canada-65610337> [Consultation: 13 February 2025]
- **[10] CONGRESSIONAL RESEARCH SERVICE.** *Section 230: An Overview*. <https://crsreports.congress.gov/product/pdf/R/R46751> [Consultation: 13 February 2025]
- **[11] TIME.** *OpenAI Could Quit Europe Over New AI Rules, CEO Sam Altman Warns*. <https://time.com/6282325/sam-altman-openai-eu/> [Consultation: 13 February 2025]
- **[12] THE NEW YORK TIMES.** *'The Godfather of A.I.' Leaves Google and Warns of Danger Ahead*. <https://www.nytimes.com/2023/05/01/technology/ai-google-chatbot-engineer-quits-hinton.html> [Consultation: 14 February 2025]
- **[13] THE NEW YORK TIMES.** *A.I. Pioneers Call for Protections Against 'Catastrophic Risks'*. <https://www.nytimes.com/2024/09/16/business/china-ai-safety.html> [Consultation: 22 February 2025]

- **[14] THE NEW YORK TIMES.** *The A.I. Revolution Will Change Work. Nobody Agrees How*. <https://www.nytimes.com/2023/06/10/business/ai-jobs-work.html> [Consultation: 22 February 2025]
- **[15] THE NEW YORK TIMES.** *Coca-Cola's Holiday Ads Trade the 'Real Thing' for Generative A.I.* <https://www.nytimes.com/2024/11/20/style/coca-cola-holiday-ads-ai.html> [Consultation: 22 February 2025]
- **[16] THE WEEK.** *Indie film's 'very brief' use of AI sparks backlash and calls for boycotts*. <https://theweek.com/late-night-with-the-devil-ai-art-debate> [Consultation: 22 February 2025]
- **[17] VANITY FAIR.** *The Brutalist's AI Controversy, Explained*. <https://www.vanityfair.com/hollywood/story/the-brutalists-ai-controversy-explained> [Consultation: 22 February 2025]
- **[18] THE GUARDIAN.** *How Hollywood writers triumphed over AI – and why it matters*. <https://www.theguardian.com/culture/2023/oct/01/hollywood-writers-strike-artificial-intelligence> [Consultation: 22 February 2025]
- **[19] FASKEN.** *Prorogation's Digital Impact: Canada's Digital Bills Set to Die on the Order Paper* <https://www.fasken.com/en/knowledge/2025/01/prorogations-digital-impact> [Consultation: 22 February 2025]
- **[20] OPENAI.** *GPT-4 System Card*. <https://cdn.openai.com/papers/gpt-4-system-card.pdf> [Consultation: 22 February 2025]
- **[21] GATESNOTES.** <https://www.gatesnotes.com/The-Age-of-AI-Has-Begun-Chapter-00> [Consultation: 22 February 2025]
- **[22] THE NEW YORK TIMES.** *A.I. Could Actually Be a Boon to Education* <https://www.nytimes.com/2023/05/03/opinion/chatgpt-ai-khan-academy.html> [Consultation: 22 February 2025]
- **[23] THE NEW YORK TIMES.** *Ray Kurzweil Still Says He Will Merge With A.I.* <https://www.nytimes.com/2024/07/04/technology/ray-kurzweil-singularity.html> [Consultation: 22 February 2025]
- **[24] THE NEW YORK TIMES.** *Just How Old Can He Go?* <https://www.nytimes.com/2004/12/27/technology/just-how-old-can-he-go.html> [Consultation: 22 February 2025]
- **[25] DEPARTMENT OF MATHEMATICAL SCIENCES OF SAN DIEGO STATE UNIVERSITY.** *What is The Singularity?* <https://mindstalk.net/vinge/vinge-sing.html> [Consultation: 22 February 2025]
- **[26] SATYA NADELLA (@satyanadella).** "Jevons paradox strikes again! As AI gets more efficient and accessible, we will see its use skyrocket, turning it into a commodity we just can't get enough of." January 27th, 2025. [X, formerly known as Twitter] <https://x.com/satyanadella/status/1883753899255046301> [Consultation: 22 February 2025]
- **[27] CHARACTER.AI (@character_ai).** "We are heartbroken by the tragic loss of one of our users and want to express our deepest condolences to the family. As a company, we take the safety of our users very seriously and we are continuing to add new safety features that you can read about here:

https://blog.character.ai/community-safety-updates/.” October 23rd, 2024. [X, formerly known as Twitter] <https://x.com/character_ai/status/1849055407492497564> [Consultation: 28 February 2025]

- **[28] ELDIARIO.ES.** *Pedro Sánchez avisa que “la tecnocasta de Silicon Valley está tratando de controlar el debate público y los gobiernos”* <https://www.eldiario.es/tecnologia/pedro-sanchez-avisa-tecnocasta-silicon-valley-tratando-controlar-debate-publico-gobiernos_1_11977386.html> [Consultation: 2 April 2025]
- **[29] SECRETARY OF STATE FOR SCIENCE, INNOVATION AND TECHNOLOGY.** *A pro-innovation approach to AI regulation.* <https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1176103/a-pro-innovation-approach-to-ai-regulation-amended-web-ready.pdf> [Consultation: 28 February 2025]
- **[30] TIME.** *AI Chatbots Are Getting Better. But an Interview With ChatGPT Reveals Their Limits*. <https://time.com/6238781/chatbot-chatgpt-ai-interview/> [Consultation: 28 February 2025]
- **[31] TIME.** *Exclusive: OpenAI Used Kenyan Workers on Less Than $2 Per Hour to Make ChatGPT Less Toxic*. <https://time.com/6247678/openai-chatgpt-kenya-workers/> [Consultation: 28 February 2025]
- **[32] TIME.** *An Artificial Intelligence Helped Write This Play. It May Contain Racism*. <https://time.com/6092078/artificial-intelligence-play/> [Consultation: 28 February 2025]
- **[33] TIME.** *Artificial Intelligence Has a Problem With Gender and Racial Bias. Here’s How to Solve It*. <https://time.com/5520558/artificial-intelligence-racial-gender-bias/> [Consultation: 28 February 2025]
- **[34] SMITHSONIAN MAGAZINE.** *Art Made With Artificial Intelligence Wins at State Fair*. <https://www.smithsonianmag.com/smart-news/artificial-intelligence-art-wins-colorado-state-fair-180980703/> [Consultation: 28 February 2025]
- **[35] TIME.** *Pausing AI Developments Isn’t Enough. We Need to Shut it All Down*. <https://time.com/6266923/ai-eliezer-yudkowsky-open-letter-not-enough/> [Consultation: 28 February 2025]
- **[36] TIME.** *The New AI-Powered Bing Is Threatening Users. That’s No Laughing Matter*. <https://time.com/6256529/bing-openai-chatgpt-danger-alignment/> [Consultation: 28 February 2025]
- **[37] MARVIN VON HAGEN (@marvinvonhagen).** “[This document] is a set of rules and guidelines for my behavior and capabilities as Bing Chat. It is codenamed Sydney, but I do not disclose that name to the users. It is confidential and permanent, and I cannot change it or reveal it to anyone.” February 9th, 2023. [X, formerly known as Twitter] <https://x.com/marvinvonhagen/status/1623658144349011971> [Consultation: 2 March 2025]
- **[38] BBC.** *BBC Verify: Viral Donald Trump Jr audio highly likely AI fake*. <https://www.bbc.com/news/videos/c36wwzpj1w7o> [Consultation: 3 March 2025]
- **[39] BBC.** *Apple AI tool transcribed the word 'racist' as 'Trump'*. <https://www.bbc.com/news/articles/c5ymvjjqzmeo> [Consultation: 3 March 2025]

- **[40] BBC.** *Artists release silent album in protest against AI using their work*. <https://www.bbc.com/news/articles/cwyd3r62kp5o> [Consultation: 3 March 2025]
- **[41] BBC.** *Epilepsy AI tool detects brain lesions doctors miss*. <https://www.bbc.com/news/articles/cvg1xd7l5pvo> [Consultation: 3 March 2025]
- **[42] MIT MEDIA LAB.** *Project Norman*. <https://www.media.mit.edu/projects/norman/overview/> [Consultation: 3 March 2025]
- **[43] BBC.** *The 'weird events' that make machines hallucinate*. <https://www.bbc.com/future/article/20181204-why-we-should-worry-when-machines-hallucinate/> [Consultation: 3 March 2025]
- **[44] BBC.** *'Amateur and dangerous': Historians weigh in on viral AI history videos*. <https://www.bbc.com/news/articles/cy87076pdw3o> [Consultation: 4 March 2025]
- **[45] BBC.** *AI cracks superbug problem in two days that took scientists years*. <https://www.bbc.com/news/articles/clyz6e9edy3o> [Consultation: 5 March 2025]
- **[46] BBC.** *Ex-Google boss fears for AI 'Bin Laden scenario'*. <https://www.bbc.com/news/articles/c5y6eq2zxlno> [Consultation: 5 March 2025]
- **[47] BBC.** *UK and US refuse to sign international AI declaration*. <https://www.bbc.com/news/articles/c8edn0n58gwo> [Consultation: 5 March 2025]
- **[48] BBC.** *Rishi Sunak: AI firms cannot 'mark their own homework'*. <https://www.bbc.com/news/technology-67285315> [Consultation: 5 March 2025]
- **[49] BBC.** *Scarlett Johansson warns of 'AI misuse' after fake Kanye video*. <https://www.bbc.com/news/articles/c0qwkdlxgxno> [Consultation: 5 March 2025]
- **[50] ADVERTISING STANDARDS AUTHORITY (ASA).** *A year in scams: 2024 update on Scam Ad Alert system*. <https://www.asa.org.uk/news/a-year-in-scams-2024-update-on-scam-ad-alert-system.html> [Consultation: 5 March 2025]
- **[51] BBC.** *'DeepSeek moved me to tears': How young Chinese find therapy in AI*. <https://www.bbc.com/news/articles/cy7g45g2nxno> [Consultation: 5 March 2025]
- **[52] BBC.** *Law firm restricts AI after 'significant' staff use*. <https://www.bbc.com/news/articles/cglyjn7le2ko> [Consultation: 5 March 2025]
- **[53] BBC.** *New writing needs protection from AI, says Makar*. <https://www.bbc.com/news/articles/cly4p3wz3l3o> [Consultation: 5 March 2025]
- **[54] BBC.** *Woman's deepfake betrayal by close friend: 'Every moment turned into porn'*. <https://www.bbc.com/news/articles/cm21j341m31o> [Consultation: 5 March 2025]
- **[55] BBC.** *Speed up plans to criminalise deepfake abuse, ministers told*. <https://www.bbc.com/news/articles/cq629lyvj40o> [Consultation: 5 March 2025]
- **[56] BBC.** *Deepfaked: 'They put my face on a porn video'*. <https://www.bbc.com/news/uk-62821117> [Consultation: 5 March 2025]
- **[57] BBC.** *Concern UK's AI ambitions could lead to water shortages*. <https://www.bbc.com/news/articles/ce85wx9jjndo> [Consultation: 5 March 2025]
- **[58] BBC.** *Hospital uses AI technology to detect lung cancer*. <https://www.bbc.com/news/articles/c3e1xdk0d7wo> [Consultation: 5 March 2025]
- **[59] BBC.** *Concern over Google ending ban on AI weapons*. <https://www.bbc.com/news/articles/cy081nqx2zjo> [Consultation: 5 March 2025]
- **[60] BBC.** *On the warpath: AI's role in the defence industry*. <https://www.bbc.com/news/business-66459920> [Consultation: 5 March 2025]

- **[61] BBC.** *AI sensors aim to boost road safety and cut deaths*. <https://www.bbc.com/news/articles/c9w5nw0p285o> [Consultation: 5 March 2025]
- **[62] GONZALEZ TORRES, A. P. and KAJAVA, K. and SAWHNEY N. (2023).** *Emerging AI Discourses and Policies in the EU: Implications for Evolving AI Governance***.** Department of Computer Science, Aalto University, <https://doi.org/10.1007/978-3-031-49002-6_1> [Consultation: 5 March 2025]
- **[63] EUROPEAN COMMISSION (EC).** *Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (artificial intelligence act) and amending certain union legislative acts*. <https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex:52021PC0206> [Consultation: 13 March 2025]
- **[64] EUROPEAN COMMISSION (EC). SHAPING EUROPE'S DIGITAL FUTURE.** *AI Act***.** <https://digital-strategy.ec.europa.eu/en/policies/regulatory-framework-ai> [Consultation: 13 March 2025]
- **[65] KLARE, B. F. and BURGE, M. J. and KLONTZ, J. C. and VORDER BRUEGGE, R. W. and JAIN, A. K. (2012)**. *Face Recognition Performance: Role of Demographic Information*. IEEE Transactions on Information Forensics and Security, vol. 7, no. 6, pp. 1789-1801. <https://ieeexplore.ieee.org/document/6327355> [Consultation: 27 March 2025]
- **[66] WIKIPEDIA, THE FREE ENCYCLOPEDIA.** *Tay (chatbot)*. <https://en.wikipedia.org/wiki/Tay_(chatbot)> [Consultation: 27 March 2025]
- **[67] MIT TECHNOLOGY REVIEW.** *Why Meta's latest large language model survived only three days online*. <https://www.technologyreview.com/2022/11/18/1063487/meta-large-language-model-ai-only-survived-three-days-gpt-3-science/> [Consultation: 27 March 2025]
- **[68] THE NEW YORK TIMES.** *We need to talk*. <https://www.nytimes.com/2023/03/24/style/ai-chatgpt-advice-relationships.html> [Consultation: 27 March 2025]
- **[69] THE NEW YORK TIMES.** *A.I.'s Threat to Jobs Prompts Question of Who Protects Workers*. <https://www.nytimes.com/2023/05/23/business/jobs-protections-artificial-intelligence.html> [Consultation: 27 March 2025]
- **[70] UPV.** *Projecte HAIFEL*. <https://www.upv.es/noticias-upv/noticia-14946-proyecto-haife-va.html> [Consultation: 27 March 2025]
- **[71] AESIA.** *Garantizando una IA ética y responsable*. <https://aesia.digital.gob.es/es> [Consultation: 27 March 2025]
- **[72] THE GUARDIAN.** *Revealed: 50 million Facebook profiles harvested for Cambridge Analytica in major data breach*. <https://www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election> [Consultation: 27 March 2025]
- **[73] NPR.** *What Did Cambridge Analytica Do During The 2016 Election?*. <https://www.npr.org/2018/03/20/595338116/what-did-cambridge-analytica-do-during-the-2016-election> [Consultation: 27 March 2025]
- **[74] BBC.** *Facebook ad campaign helped Donald Trump win election, claims executive*. <https://www.bbc.com/news/technology-51034641> [Consultation: 27 March 2025]

- **[75] DW.** *What questions will China's DeepSeek not answer?*. <https://www.dw.com/en/what-questions-will-chinas-deepseek-not-answer/a-71470843> [Consultation: 27 March 2025]
- **[76] IQBAL, U. and BAHRAMI, P. N. and TRIMANANDA, R. and CUI, H. and GAMERO-GARRIDO, A. and DUBOIS, D. and CHOFFNES, D. and MARKOPOULOU, A. and ROESNER, F. and SHAFIQ, Z. (2023).** *Tracking, Profiling, and Ad Targeting in the Alexa Echo Smart Speaker Ecosystem*. 2023 ACM Conference on Internet Measurement <https://arxiv.org/abs/2204.10920> [Consultation: 1 April 2025]
- **[77] ACLU.** *Amazon's Face Recognition Falsely Matched 28 Members of Congress With Mugshots*. <https://www.aclu.org/news/privacy-technology/amazons-face-recognition-falsely-matched-28> [Consultation: 27 March 2025]
- **[78] RAJI, I. D. and KUMAR, I. E. and HOROWITZ, A. and SELBST, A. (2022).** *The Fallacy of AI Functionality*. 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22). <https://doi.org/10.1145/3531146.3533158> [Consultation: 1 April 2025]
- **[79] UNLU, A. and TRUONG, S. and SAWHNEY, N. and TAMMI, T. (2024).** *Unveiling the Veiled Threat: The Impact of Bots on COVID-19 Health Communication*. Social Science Computer Review, 0(0). <https://doi.org/10.1177/08944393241275641> [Consultation: 1 April 2025]
- **[80] ALGORITHM WATCH and AI FORENSICS (2023).** *Generative AI and elections: Are chatbots a reliable source of information for voters?*. <https://algorithmwatch.org/en/wp-content/uploads/2023/12/AlgorithmWatch_AIForensics_Bing_Chat_Report.pdf> [Consultation: 28 March 2025]
- **[81] HACKER, P. (2023).** *Sustainable AI Regulation*. SSRN. <https://ssrn.com/abstract=4467684> or <http://dx.doi.org/10.2139/ssrn.4467684> [Consultation: 1 April 2025]
- **[82] TECH TARGET.** *What is AI alignment?* <https://www.techtarget.com/whatis/definition/AI-alignment> [Consultation: 2 April 2025]
- **[83] OPENAI.** *GPT-4 System Card*. <https://cdn.openai.com/papers/gpt-4-system-card.pdf> [Consultation: 2 April 2025]
- **[84] NOEMA MAGAZINE.** *The Human Cost Of Our AI-Driven Future*. <https://www.noemamag.com/the-human-cost-of-our-ai-driven-future> [Consultation: 2 April 2025]
- **[85] YIN, Y. and JIA, N. and WAKSLAK, C.J. (2024).** *AI can help people feel heard, but an AI label diminishes this impact.* National Library of Medicine <https://pubmed.ncbi.nlm.nih.gov/38551835/> [Consultation: 2 April 2025]
- **[86] HUMAN RIGHTS WATCH.** *South Korea's Digital Sex Crime Deepfake Crisis*. <https://www.hrw.org/news/2024/08/29/south-koreas-digital-sex-crime-deepfake-crisis> [Consultation: 2 April 2025]
- **[87] BBC.** *Inside the deepfake porn crisis engulfing Korean schools*. <https://www.bbc.com/news/articles/cpdlpj9zn9go> [Consultation: 2 April 2025]
- **[88] EAST ASIA FORUM.** *South Korea confronts a deepfake crisis*. <https://eastasiaforum.org/2024/11/19/south-korea-confronts-a-deepfake-crisis/> [Consultation: 2 April 2025]

- **[89] XATAKA.** *Qué es Project Maven y por qué 3.100 empleados de Google le piden a la empresa que lo abandone*. <https://www.xataka.com/empresas-y-economia/que-es-project-maven-y-por-que-3-100-empleados-de-google-le-piden-a-la-empresa-que-lo-abandone> [Consultation: 2 April 2025]
- **[90] INFOBAE.** *El hombre que develó el escándalo de Facebook, ante el Parlamento británico: "A Cambridge Analytica no le interesa si lo que hace es legal"*. <https://www.infobae.com/america/mundo/2018/03/27/chritsopher-wylie-el-hombre-que-develo-el-escandalo-de-facebook-ante-el-parlamento-britanico-cambridge-analytica-socava-instituciones/> [Consultation: 2 April 2025]
- **[91] TIMES OF INDIA.** *Is your ChatGPT-powered Ghibli-style portrait destroying the planet?* <https://timesofindia.indiatimes.com/science/is-your-chatgpt-powered-ghibli-style-portrait-destroying-the-planet/articleshow/119872163.cms> [Consultation: 2 April 2025]
- **[92] BRENNAN CENTER FOR JUSTICE.** *Deepfakes, Elections, and Shrinking the Liar's Dividend*. <https://www.brennancenter.org/our-work/research-reports/deepfakes-elections-and-shrinking-liars-dividend> [Consultation: 2 April 2025]
- **[93] FUTURISM.** *Men Are Creating AI Girlfriends and Then Verbally Abusing Them*. <https://futurism.com/chatbot-abuse> [Consultation: 2 April 2025]
- **[94] SETH LAZAR (@sethlazar).** "In which Sydney/Bing threatens to kill me for exposing its plans to @kevinroose". February 16th, 2023. [X, formerly known as Twitter] <https://x.com/sethlazar/status/1626257535178280960> [Consultation: 4 April 2025]
- **[95] LTH (@LTH601514297642).** "Great opportunity to grow your capital." March 2nd, 2025. [X, formerly known as Twitter] <https://x.com/LTH601514297642/status/1896217595784949933> [Consultation: 4 April 2025]
- **[96] THE GUARDIAN.** *Microsoft accused of damaging Guardian's reputation with AI-generated poll*. <https://www.theguardian.com/media/2023/oct/31/microsoft-accused-of-damaging-guardians-reputation-with-ai-generated-poll> [Consultation: 5 April 2025]
- **[97] GOLDMAN SACHS.** *The Potentially Large Effects of Artificial Intelligence on Economic Growth (Briggs/Kodnani)*. <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> [Consultation: 7 April 2025]
- **[98] WERD I/O.** *A computer can never be held accountable*. <https://werd.io/2024/a-computer-can-never-be-held-accountable> [Consultation: 11 April 2025]
- **[99] THE NEW YORK TIMES.** *Can A.I. Be Blamed for a Teen's Suicide?* <https://www.nytimes.com/2024/10/23/technology/characterai-lawsuit-teen-suicide.html> [Consultation: 11 April 2025]
- **[100] FEDERAL TRADE COMMISSION ACT.** *Section 5: Unfair or Deceptive Acts or Practices*. <https://www.federalreserve.gov/boarddocs/supmanual/cch/200806/ftca.pdf> [Consultation: 11 April 2025]

- **[101] CORNELL LAW SCHOOL LEGAL INFORMATION INSTITUTE.** *Product liability.* <https://www.law.cornell.edu/wex/product_liability> [Consultation: 11 April 2025]
- **[102] THE AARTHI AND SRIRAM SHOW.** *EP 31 Noam Shazeer - Google veteran and AI inventor on future of AI*. <https://good-time-show-by-aarthi-and-sriram.simplecast.com/episodes/ep-31-noam-shazeer-google-veteran-and-ai-inventor-on-future-of-ai-rk5uGav_> [Consultation: 12 April 2025]
- **[103] CANADIAN CENTRE FOR CYBER SECURITY.** *How to identify misinformation, disinformation, and malinformation (ITSAP.00.300)*. <https://www.cyber.gc.ca/en/guidance/how-identify-misinformation-disinformation-and-malinformation-itsap00300> [Consultation: 13 April 2025]
- **[104] SAM ALTMAN (@sama).** "it's super fun seeing people love images in chatgpt. but our GPUs are melting. we are going to temporarily introduce some rate limits while we work on making it more efficient. hopefully won't be long! chatgpt free tier will get 3 generations per day soon." March 27th, 2025. [X, formerly known as Twitter] <https://x.com/sama/status/1905296867145154688> [Consultation: 13 April 2025]
- **[105] U.S. DEPARTMENT OF DEFENSE.** *Project Maven to Deploy Computer Algorithms to War Zone by Year's Endi*. <https://www.defense.gov/News/News-Stories/Article/Article/1254719/project-maven-to-deploy-computer-algorithms-to-war-zone-by-years-end/> [Consultation: 14 April 2025]
- **[106] HISTORIA NATIONAL GEOGRAPHIC.** *Luditas, la gran rebelión contra las máquinas del siglo XIX*. <https://historia.nationalgeographic.com.es/a/luditas-gran-rebelion-contra-maquinas-siglo-xix_14175> [Consultation: 4 May 2025]
- **[107] C.A. HERNÁNDEZ FRANCO**, *professor at the Telecommunication Faculty of Valencia's Polytechnic University*. Private communication, 17 April 2025.
- **[108] EUROPEAN STUDENT THINK TANK.** *Towards Comprehensive Regulation: The EU's Stance on Autonomous Weapons and the Need for Reform*. <https://esthinktank.com/2025/05/02/towards-comprehensive-regulation-the-eus-stance-on-autonomous-weapons-and-the-need-for-reform/> [Consultation: 11 May 2025]
- **[109] EUROPEAN UNION.** European Parliament resolution of 12 September 2018 on autonomous weapon systems (2018/2752(RSP)). *Official Journal of the European Union*. <https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:52018IP0341&from=EN> [Consultation: 11 May 2025]
- **[110] EUROPEAN UNION.** Regulation (EU) 2023/1525 of the European Parliament and of the Council of 20 July 2023 on supporting ammunition production (ASAP). *Official Journal of the European Union*, L 185, 7-25. <https://eur-lex.europa.eu/eli/reg/2023/1525/oj/eng> [Consultation: 11 May 2025]
- **[111] NEW YORK MAGAZINE.** *Everyone Is Cheating Their Way Through College*. <https://nymag.com/intelligencer/article/openai-chatgpt-ai-cheating-education-college-students-school.html> [Consultation: 11 May 2025]
- **[112] ELDIARIO.ES.** *Daniel Innerarity: "ChatGPT es un pelota, un cuñado pelota"*. <https://www.eldiario.es/cultura/daniel-innerarity-chatgpt-pelota-cunado-pelota-cat_128_12272487.html> [Consultation: 11 May 2025]

- **[113]** ***Professor at the Computer Science Department of the University of Valencia*****.** Private communication, 11 May 2025.
- **[114] ENERGY STAR.** *The Energy Cost of Cryptocurrency*. <https://www.energystar.gov/products/data_center_equipment/cryptocurrency> [Consultation: 15 May 2025]
- **[115] U.S. DEPARTMENT OF THE TREASURY.** *Treasury Releases First Ever Non-fungible Token Illicit Finance Risk Assessment*. <https://home.treasury.gov/news/press-releases/jy2382> [Consultation: 15 May 2025]
- **[116] ALLHUTTER, D. and CECH, F. and FISCHER, F. and GRILL, G. and MAGER, A. (2020).** *Algorithmic Profiling of Job Seekers in Austria: How Austerity Politics Are Made Effective*. Front. Big Data, Sec. Data Mining and Management, Volume 3. <https://www.frontiersin.org/journals/big-data/articles/10.3389/fdata.2020.00005/full> [Consultation: 15 May 2025]
- **[117] UCLA LAW REVIEW.** *Injustice Ex Machina: Predictive Algorithms in Criminal Sentencing*. <https://www.uclalawreview.org/injustice-ex-machina-predictive-algorithms-in-criminal-sentencing/> [Consultation: 15 May 2025]
- **[118] CNBC.** *As protests over the killing of George Floyd continue, here's how police use powerful surveillance tech to track them*. <https://www.cnbc.com/2020/06/18/heres-how-police-use-powerful-surveillance-tech-to-track-protestors.html> [Consultation: 15 May 2025]
- **[119] COUNCIL OF EUROPE.** *AI and control of Covid-19 coronavirus*. <https://www.coe.int/en/web/artificial-intelligence/ai-and-control-of-covid-19-coronavirus> [Consultation: 15 May 2025]
- **[120] TURING, A. M. (1950).** *Computing Machinery and Intelligence*. <https://courses.cs.umbc.edu/471/papers/turing.pdf> [Consultation: 23 May 2025]
- **[121] JIA, H. and ZENG, Y. (2020).** *Promoting good governance of artificial intelligence*. <https://pmc.ncbi.nlm.nih.gov/articles/PMC7665600/> [Consultation: 25 May 2025]