ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

ΤΜΗΜΑ ΨΗΦΙΑΚΩΝ ΣΥΣΤΗΜΑΤΩΝ

Πρόγραμμα Μεταπτυχιακών Σπουδών

«ΔΙΚΑΙΟ ΚΑΙ ΤΕΧΝΟΛΟΓΙΕΣ ΠΛΗΡΟΦΟΡΙΚΗΣ ΚΑΙ ΕΠΙΚΟΙΝΩΝΙΩΝ»

Ακαδημαϊκό έτος 2024-2025

ΜΕΤΑΠΤΥΧΙΑΚΗ ΔΙΠΛΩΜΑΤΙΚΗ ΕΡΓΑΣΙΑ

του ΧΡΙΣΤΟΥ ΚΟΥΡΟΥΚΛΗ (Α.Μ.: ΜΔΙ2329)

ΥΠΕΥΘΥΝΗ ΤΕΧΝΗΤΗ ΝΟΗΜΟΣΥΝΗ: ΕΠΑΝΑΚΑΘΟΡΙΖΟΝΤΑΣ ΤΑ ΟΡΙΑ ΤΗΣ ΡΥΘΜΙΣΗΣ ΤΗΣ ΤΕΧΝΟΛΟΓΙΑΣ

RESPONSIBLE AI: REDEFINING THE BOUNDARIES OF TECHNOLOGY REGULATION

Επιβλέπουσα:

ΛΙΛΙΑΝ ΜΗΤΡΟΥ

Πειραιάς, Μάρτιος, 2025

TABLE OF CONTENTS

SUMMARY

AI is a driving force in today's world. Its immense capabilities, however, come with a myriad risks. Automated weapons, privacy infringement, biometric surveillance systems, workforce replacement, to name but a few, present grave threats for humanity's core values. To minimize those risks, experts have set apart certain principles that AI systems shall uphold. Transparency, robustness, safety, data governance and human oversight are among the most commonly sighted of these principles. AI systems with such characteristics could earn the characterization of Responsible AI systems (RAI).

To enforce these principles, the European Union (EU) has voted into effect the AI Act (AIA). This Regulation categorizes AI systems according to the risks they pose. Systems that are deemed to pose unacceptable risks are banned from being put into service altogether. High-risk systems must comply with extensive requirements such as human oversight, transparency prerequisites, built-in fail-safes, log keeping and many more. From a technical standpoint, the International Organization for Standardization (ISO) has issued two ISO standards related to AI safety, while the National Institute of Standards and Technology has published its own AI risk management playbook.

AIA is a truly ambitious piece of legislation. Attempting to regulate what might prove to be the final technological frontier. However, its importance is greatly undermined by the lack of the EU's AI development programs to date. Most of the AI-related research and development is undertaken by US-based and Chinese companies. Their attitude towards the compliance requirements set by the AI Act is mainly focused on ignoring them.

Moreover, both the USA and the PRC seem more focused on becoming the leading AI force than becoming the leading force in Responsible AI. This attitude is evident in the latest decisions made by the US Administration, and it is rapidly gaining ground amongst the leading AI-related companies as well. Fuelled by China's latest advancements in the sector, the USA's reaction tends to establish a cold war-like arms race environment. This climate isn't kind to regulation attempts. In fact, the latest trend in the USA is to deregulate AI research, in an attempt to maximize the sector's agility. The allocation of astronomical funds for AI research coupled with a sense of combative urgency constitute the perfect prerequisites for the creation of AI systems that will in no way be responsible.

The ineffectiveness of regional legislation such as AIA, combined with the international power dynamics developed, underline the importance for the creation of an international AI governance framework. Such a framework could lead to the adoption of common AI safety practices, slowing down the pace of innovation, defusing the arms race dynamic and hopefully better gatekeep the goal of Responsible AI development.

This plan isn't without soft spots. Its implementation will require great political will from all the countries involved, an ingredient that, at the moment, seems to be missing. In addition, developments during this millennium have brought similar international initiatives under question. The outbreak of major wars and military conflicts, global public health breakdowns and the management of the existential climate crisis have cast doubt on the effectiveness of international organizations and multi-national cooperation schemes. Last but not least, the question of the technical aspect of AI regulation remains. Many of the challenges presented by AI, such as the alignment and control problems, the issues of opacity and untraceability, the demand for explainable systems, subject to human oversight, might not even be sustainably technically possible, despite the most robust international Regulations we can vote into effect and follow religiously.

ΠΕΡΙΛΗΨΗ

Η τεχνητή νοημοσύνη (ΤΝ) έχει ραγδαίως αναδειχθεί σε κινητήρια δύναμη της εποχής μας. Οι τεράστιες δυνατότητές της συνοδεύονται από πληθώρα κινδύνων. Η χρήση αυτοματοποιημένων όπλων και συστημάτων ΤΝ για την εξαπόλυση κυβερνοεπιθέσεων, η εγκατάσταση συστημάτων βιομετρικής παρακολούθησης, οι κίνδυνοι που παρουσιάζει για την ιδιωτικότητα και η ραγδαία αναδιαμόρφωση της αγοράς εργασίας που αναμένεται να σημειωθεί εξαιτίας της ΤΝ, αποτελούν μόνο κάποια από τα ζητήματα που χρίζουν εξέτασης.

Για την ελαχιστοποίηση αυτών των κινδύνων, οι ειδικοί έχουν ξεχωρίσει συγκεκριμένες αξίες που πρέπει να υποστηρίζουν τα συστήματα ΤΝ. Διαφάνεια, ανθεκτικότητα, ασφάλεια, διαχείριση δεδομένων και ανθρώπινη εποπτεία είναι μεταξύ των πιο συχνά αναφερόμενων αξιών. Συστήματα ΤΝ με τέτοια χαρακτηριστικά θα μπορούσαν να χαρακτηριστούν ως συστήματα υπεύθυνης ΤΝ.

Για την επικράτηση των αρχών αυτών και την επίτευξη της δημιουργίας συστημάτων υπεύθυνης ΤΝ, η ΕΕ έχει ψηφίσει και θέσει σε ισχύ τον Κανονισμό για την Τεχνητή Νοημοσύνη («ο Κανονισμός»). Αυτός ο Κανονισμός κατηγοριοποιεί τα συστήματα ΤΝ σύμφωνα με τους κινδύνους που ενέχουν. Συστήματα που κρίνεται ότι παρουσιάζουν απαγορευτικούς κινδύνους αποκλείονται ολοσχερώς από την ενωσιακή αγορά. Τέτοια συστήματα είναι αυτά που ταξινομούν πολίτες σε ομάδες βάσει κοινωνικής βαθμολογίας, που επιτρέπουν την εξ αποστάσεως βιομετρική ταυτοποίηση προσώπων και μία σειρά άλλων συστημάτων τα οποία απαριθμούνται στο άρθρο 5 του Κανονισμού.

Ακολουθούν τα συστήματα υψηλού κινδύνου τα οποία πρέπει να συμμορφώνονται με εκτεταμένες απαιτήσεις. Τέτοιες απαιτήσεις έχουν να κάνουν με την ενσωμάτωση συστημάτων διαχείρισης κινδύνων, την αυστηρή διακυβέρνηση των δεδομένων που αξιοποιούνται από τα συστήματα ΤΝ, τη σύνταξη και την επικαιροποίηση τεχνικού φακέλου, ο οποίος πρέπει να περιλαμβάνει πλήθος πληροφοριών αναφορικά με το σύστημα, την αυτοματοποιημένη δημιουργία αρχείου δραστηριότητας του συστήματος και τη σχεδίαση του κατά τρόπου που θα επιτρέπει την εποπτεία του, την αποσαφήνιση των σταδίων που μεσολάβησαν μεταξύ της τροφοδοσίας του με κάποια δεδομένα και της παραγωγής ενός συγκεκριμένου αποτελέσματος, ενώ ταυτόχρονα θα είναι ικανό να αποκρούσει τυχόν κυβερνοεπιθέσεις.

Λιγότερες υποχρεώσεις, κυρίως διαφάνειας θεσπίζονται για τα συστήματα ΤΝ χαμηλού κινδύνου. Επιπλέον ο Κανονισμός προβλέπει την ίδρυση και λειτουργιά ενός πολύπλοκου συστήματος διακυβέρνησης σε ενωσιακό επίπεδο αποτελούμενο από ευρωπαϊκές αρχές όπως η Υπηρεσία ΤΝ και το Συμβούλιο Τεχνητής Νοημοσύνης αλλά και εθνικές επιβλέπουσες αρχές. Ένας άλλος θεσμός με εξαιρετικό ενδιαφέρον ο οποίος θεσπίζεται δια του Κανονισμού είναι τα ρυθμιστικά δοκιμαστήρια ΤΝ τα οποία αποσκοπούν στην πιστοποίηση της ασφάλειας ενός συστήματος προτού αυτό τεθεί σε λειτουργιά στην ενωσιακή αγορά.

Λιγότερο συγκροτημένες είναι οι αντίστοιχες προσπάθειες που έχουν λάβει χώρα στις ΗΠΑ. Οι σημαντικότερες προσπάθειες εντοπίζονται σε πολιτειακό επίπεδο, ενώ η ομοσπονδιακή κυβέρνηση είχε αρχικώς κάνει κάποιες προσπάθειες να υιοθετήσει αντίστοιχα του Κανονισμού νομοθετήματα, οι οποίες ωστόσο, με την επανεκλογή του Προέδρου Τραμπ έχουν διακοπεί. Αντιθέτως, η νέα κυβέρνηση φαίνεται να προσανατολίζεται προς την εφαρμογή λιγότερων και όχι περισσότερων υποχρεώσεων για τους δημιουργούς συστημάτων ΤΝ.

Σε επίπεδο τεχνικών προδιαγραφών, ο Διεθνής Οργανισμός Τυποποίησης (ISO) έχει εκδώσει δύο πρότυπα για την ασφάλεια της ΤΝ, ενώ το Εθνικό Ινστιτούτο Προτύπων και Τεχνολογίας έχει εκδώσει το δικό του πλαίσιο διαχείρισης κινδύνων ΤΝ.

Ο Κανονισμός για την ΤΝ είναι ένα πραγματικά φιλόδοξο νομοθέτημα. Επιχειρεί να ρυθμίσει ένα από τα πλέον αντιδραστικά στη ρύθμιση πεδία. Ωστόσο, η σημασία του υπονομεύεται σημαντικά από την έλλειψη προγραμμάτων ανάπτυξης ΤΝ της ΕΕ μέχρι σήμερα. Μόλις τον Φεβρουάριου του 2025 η Πρόεδρος της Ευρωπαϊκής Επιτροπής ανακοίνωσε σημαντικά προγράμματα επένδυσης στην ανάπτυξη συστημάτων ΤΝ. Το μεγαλύτερο μέρος της έρευνας και ανάπτυξης σχετικά με την ΤΝ εκτελείται από εταιρείες με έδρα τις ΗΠΑ και την Κίνα. Η στάση τους απέναντι στις απαιτήσεις συμμόρφωσης που θέτει ο Νόμος για την ΤΝ επικεντρώνεται κυρίως στο να τις αγνοούν, όπως προκύπτει και από δηλώσεις εξεχόντων του χώρου προσώπων όπως ο Σαμ Άλτμαν.

Επιπλέον, τόσο οι ΗΠΑ όσο και η Κίνα φαίνεται να επικεντρώνονται περισσότερο στο να εδραιωθούν ως η κυρίαρχη δύναμη στην ΤΝ παρά στο να ποδηγετήσουν τις προσπάθειες για δημιουργία υπεύθυνης ΤΝ. Αυτή η στάση είναι εμφανής στις πρόσφατες αποφάσεις της αμερικανικής κυβέρνησης, και κερδίζει ραγδαία έδαφος και μεταξύ των κορυφαίων

εταιρειών που σχετίζονται με την ΤΝ. Τροφοδοτούμενες από τις τελευταίες εξελίξεις στην Κίνα με την λειτουργία του σαφώς φθηνότερου από τα αμερικάνικα συστήματα δημιουργικής ΤΝ (ChatGPT, Claude AI), όμως εξίσου αποτελεσματικού με αυτά, DeepSeek, οι αντιδράσεις των ΗΠΑ τείνουν να εδραιώσουν ένα περιβάλλον ανταγωνισμού που ομοιάζει με τον ψυχροπολεμικό ανταγωνισμό εξοπλισμού με πυρηνικά όπλα. Αυτό το κλίμα δεν είναι πρόσφορο για την ευδοκίμηση των προσπαθειών ρύθμισης της ΤΝ. Στην πραγματικότητα, η τελευταία τάση στις ΗΠΑ είναι η απορρύθμιση της έρευνας στην ΤΝ, σε μια προσπάθεια να μεγιστοποιηθεί η ευελιξία του τομέα. Η διάθεση αστρονομικών κονδυλίων για την έρευνα στην ΤΝ, σε συνδυασμό με μια αίσθηση κατεπείγοντος αποτελούν τις τέλειες προϋποθέσεις για τη δημιουργία συστημάτων ΤΝ που σε καμία περίπτωση δεν θα μπορούν να χαρακτηριστούν υπεύθυνα.

Η αναποτελεσματικότητα της τοπικής νομοθεσίας όπως αυτή του ευρωπαϊκού Κανονισμού για την Τεχνητή Νοημοσύνη, σε συνδυασμό με τις διεθνείς δυναμικές που αναπτύσσονται υπογραμμίζουν τη σημασία της δημιουργίας ενός διεθνούς πλαισίου διακυβέρνησης για την ΤΝ. Ένα τέτοιο πλαίσιο θα μπορούσε να οδηγήσει στην υιοθέτηση κοινών πρακτικών ασφάλειας για την ΤΝ, επιβραδύνοντας τον ρυθμό καινοτομίας, αποκλιμακώνοντας τη δυναμική του ανταγωνισμού εξοπλισμών, με την ελπίδα να διαφυλάξει καλύτερα τον στόχο της υπεύθυνης ανάπτυξης ΤΝ. Την ίδια γνώμη έχει εκφράσει και ο Οργανισμός Ηνωμένων Εθνών και οι G7.

Αυτό το σχέδιο δεν στερείται αδύναμων σημείων. Η εφαρμογή του θα απαιτήσει επίδειξη πολιτικής βούλησης από όλες τις εμπλεκόμενες τεχνολογικές δυνάμεις, ένα στοιχείο που, αυτή τη στιγμή, φαίνεται να λείπει. Επιπλέον, οι εξελίξεις κατά τη διάρκεια αυτής της χιλιετίας έχουν θέσει υπό αμφισβήτηση παρόμοιες διεθνείς πρωτοβουλίες. Το ξέσπασμα πολέμων και στρατιωτικών συγκρούσεων, οι παγκόσμιες κρίσεις δημόσιας υγείας και η διαχείριση της κλιματικής κρίσης έχουν δημιουργήσει αμφιβολίες για την αποτελεσματικότητα των διεθνών οργανισμών και των διεθνικών συνεργασιών. Τέλος, παραμένει το ζήτημα της τεχνικής πτυχής της ρύθμισης της ΤΝ. Πολλές από τις προκλήσεις που παρουσιάζει η ΤΝ, όπως τα προβλήματα ευθυγράμμισης και ελέγχου, τα ζητήματα αδιαφάνειας και οι απαιτήσεις για επεξηγήσιμα συστήματα που υπόκεινται σε ανθρώπινη εποπτεία, μπορεί να μην είναι τεχνικά εφικτά, παρά τους πιο ισχυρούς διεθνείς Κανονισμούς που μπορούμε να ψηφίσουμε και να ακολουθήσουμε πιστά.

1. INTRODUCTION

It was the summer of 1956, at Dartmouth College, when a group of 10 scientists kicked off research in the field of Artificial Intelligence (AI for short), pioneered by computer scientist John McCarthy, credited with the term Artificial Intelligence. Of course, the idea of AI was not new. As early as 1950, the father of computer science, Alan Turing, wrote that a computer will be worthy of the description "intelligent" when it succeeds in deceiving a human being by convincing him that it is a human being itself. In the years that followed, the pioneers of AI were faced with colossal obstacles, such as the lack of sufficient datasets and the use of hardware with limited dynamics (Bostrom, 2014, p.6). However, technological advances have freed science from such limitations and brought a new boom in AI research.

The first big moment for AI researchers came on February 10, 1996, when IBM's Deep Blue computer beat Garry Kasparov in chess. The world champion claimed to have witnessed a demonstration of real intelligence and original thinking by his opponent (Bostrom, 2014, p.16) The day marked the end of the long-held theory of AI researchers that " "if one could devise a successful chess machine, one would seem to have penetrated to the core of human intellectual endeavour". Inscribed in Deep Blue's code, however, was a wealth of information about strategies in chess.

The next AI victory against humans came in the ancient Chinese game of Go. DeepMind's AlphaGo program, which on March 9, 2016, first beat grandmaster Lee Sedol by shocking him with its original tactics, had learned to play Go on its own, through a method known as deep reinforcement learning, using neural networks. Lee Sedol himself later stated that he felt powerless against the AI model (Harari, 2017, p.318). AlphaGo was the true forerunner of what was to come. However, even then, the term AI didn't invade our everyday lives, solidifying it as the technology of the future in the minds of people everywhere.

This major change in the public's perception of AI was marked by the release of ChatGPT in late 2022 (Bubeck, 2023, p.35). This model, developed by OpenAI, is a large language model (LLM) trained on a huge amount of text and images. Its ability to simulate human conversations and produce human-like text has led to excitement but also concerns about its potential impact. ChatGPT managed to spark a public debate about the potential and risks of generative AI,

something no other AI system had previously achieved. Indeed, the release of ChatGPT and the next versions shifted attention from Web3 technologies to generative AI. ChatGPT is considered a milestone in the history of AI, as it is the first system that can interact with users in a truly useful way. Countless AI models with similar capabilities have followed suit, such as Claude AI, DeepSeek, Notebook LM, to name a few. Post ChatGPT it feels like this time AI is here to stay.

Large language models, like the aforementioned, are not the only applications AI can have. More worrying, for example, is the expanded use of AI programmes for the automated assessment of potential employees (Matsumi, Solove, 2024, p.17), borrowers or foster parents, applicants for state benefits or tissue and organ transplants. Or the development of weapon systems and surveillance systems which, using AI programmes, will process data, predict human behaviour, collect information, identify and track potential targets and ultimately make autonomous decisions to eliminate or not eliminate their targets without any human intervention. Such applications have led many techno-sceptics to oppose the development of AI in general, while thousands of pro-AI researchers and scientists constantly highlight the potential risks. Nick Bostrom has described AI as an existential danger for humanity.

The rapid advancement and integration of artificial intelligence (AI) across society has sparked intense debate about its governance and ethical implications. As AI systems increasingly influence critical decisions in healthcare, finance, criminal justice, and other domains, the need for Responsible Artificial Intelligence (RAI) frameworks has become paramount. RAI refers to the development and implementation of AI technologies that adhere to principles of fairness, transparency, accountability, and privacy protection. However, competing narratives about AI's future impact complicate the path toward effective regulation.

The discourse surrounding AI development encompasses diverse perspectives. Techno-optimists advocate for AI as a "force for good" that merely requires appropriate guardrails, while techno-doomers warn of potential existential risks to humanity. Meanwhile, sceptics argue that current AI systems, despite their limitations, have already achieved concerning levels of societal influence through commercial exploitation and surveillance capabilities. These divergent viewpoints significantly influence policymakers' mental models and, consequently, the regulatory frameworks being developed.

Several critical challenges characterize the current AI landscape. Commercial actors' relentless pursuit of data collection and exploitation raises serious privacy concerns, while the gradual

normalization of invasive practices threatens to erode social boundaries. Labor market disruption, the spread of disinformation, and the potential for manipulation present additional challenges. Recent controversies involving prominent AI researchers and industry leaders have highlighted these issues, with figures like Geoffrey Hinton[1] raising alarms about AI's trajectory while companies simultaneously accelerate development (Brown, 2023).

In response to these challenges, various regulatory initiatives have emerged globally. The European Union's Artificial Intelligence Act[2] represents the most comprehensive attempt at AI regulation to date, while the United States has proposed an AI Bill of Rights[3]. These efforts generally emphasize key principles including fairness, transparency, accountability, privacy, and safety. However, implementing these principles requires moving beyond procedural strategies to more substantive approaches.

This paper examines the evolving landscape of Responsible AI, analysing the effectiveness of current regulatory frameworks, looking into the proposed solutions. By evaluating the interplay between commercial imperatives and ethical considerations, we aim to contribute to the development of more robust and effective AI governance structures. The analysis will focus particularly on how regulatory frameworks try to balance innovation with ethical accountability while addressing the fundamental challenges of data exploitation, algorithmic bias, and the need for transparent decision-making processes in AI systems, commenting on the current state of AI research and regulation and will attempt to evaluate their effectiveness in bringing Responsible AI systems to life.

2. AI RISKS OVERVIEW

Before delving into what Responsible AI is, what we are doing to achieve and examining if such a goal is attainable, we shall briefly consider the risks AI poses to further understand the importance of Responsible AI.

2.1 **Bias and Information Hazards**

[1] Why neural net pioneer Geoffrey Hinton is sounding the alarm on AI | MIT Sloan
[2] Regulation - EU - 2024/1689 - EN - EUR-Lex
[3] Blueprint for an AI Bill of Rights | OSTP | The White House

Certain AI-related risks stem from AI's technical characteristics - limitations. Algorithmic bias is one of these limitations presenting substantial risk. Such bias constitutes a form of procedural discrimination that may violate equal protection principles under various legal and ethical sets of rules. The accuracy deficiencies exhibited by generative AI systems further present significant evidentiary reliability concerns in contexts where decisional integrity is legally mandated, as in courtroom decisions (Coglianese, Ben Dor, 2021, p.804)

2.2 **Human Exploitation Risks**

Furthermore, AI, as many other scientific developments, begs the question of deployment. Malicious deployment in the form of deliberate weaponization of AI systems to produce falsified media (deepfake technology) and coordinate disinformation campaigns represents a clear and present danger to public discourse integrity (United Nations, 2023, p.8). From a legal standpoint, such activities can potentially be grounds for serious penalties and, depending on jurisdiction, may violate emerging synthetic media disclosure statutes and election interference prohibitions.

Another prominent risk is that of actual weaponization of AI and the creation of sophisticated autonomous weapons[4]. Usage of drones, or, more eloquently, unmanned aerial vehicles, is already on the rise, becoming more and more useful in military confrontations across the globe. However, an even bigger danger may be posed by a much less visually striking AI military application. That of AI cyberattack systems[5]. Countries possessing advanced AI cyberweapons might soon be able to cripple critical infrastructure of entire countries from their military operations' room. The proliferation of autonomous weapon systems constitutes a critical area requiring international legal scrutiny. Under the international laws of armed conflict and international humanitarian law principles, weapons systems lacking meaningful human control may fundamentally violate distinction and proportionality requirements (Cummings,2018, p.7). The emergence of an AI arms race presents additional challenges to strategic stability frameworks as the strategic advantage they'll provide is nuclear like.

Another commonly examined risk orbits around the protection of privacy and personal data. The deployment of real-time biometric surveillance systems raises profound constitutional and

[4] topkillerrobots.org/stop-killer-robots/facts-about-autonomous-weapons/
[5] https://www.iiss.org/publications/strategic-comments/2019/artificial-intelligence-and-offensive-cyber-weapons/

human rights concerns regarding reasonable expectations of privacy and the potential for chilling effects on protected expression[6]. As we will examine, jurisdictions have already correctly identified such applications as presenting unacceptable risks and taken measures to ban their usage altogether.

2.3 **Human-Machine Interaction Risks**

A less profound AI risk category arises when one examines how AI will affect our everyday life. High levels of automation, achieved with the help of AI systems can lead to substantial workforce displacement (Bubeck, 2023, p.90) marking a shift in the labour's market characteristics, surprising that caused by the Industrial Revolution, leading to unfathomable mass unemployment (United Nations, 2023, p.8).

But it is not only our work life that might be about to drastically change. The increasing AI mediation of interpersonal interactions presents novel regulatory challenges concerning fiduciary relationships and duty of care standards that current legal frameworks may be inadequate to address (Card, 1983).

2.4 **Containment and Alignment considerations**

The notion of an AI system running wild, or maybe running extremely by the book, has long been one of AI researchers' main topics of discussion, many of them, considering it the ultimate existential threat presented by AI technologies. One consideration is that an AI system might undermine human control, disabling any fail safes that might have been put in place, much like HAL 9000, the fictional AI character in the visionary 1968 film 2001: A Space Odyssey, causing havoc. Another consideration calls for a resolution to the AI Alignment problem. The solution requires aligning an AI system's goals and values to humanity's (Kökciyan,2022, p.1403). A famous thought experiment to illustrate the alignment problem is provided by Nick Bostrom, known as the paperclip maximiser.

[6] According to Cathy O'Neil «Officials in Boston, for example, were considering using security cameras to scan thousands of faces at outdoor concerts. This data would be uploaded to a service that could match each face against a million others per second». (2016, p.72)

It starts by requesting us to imagine a superintelligent AI system put into effect with the simple goal of making paperclips. The system might develop unexpected strategies to achieve its goal. Namely, the single-minded pursuit of its goal might lead to the consumption of all available resources in order to maximize paperclip production. To achieve its goal, it will be incentivised to oppose any attempts to turn it off, as once turned off, it won't be able to achieve its goal. Soon, all of the known universe, humans included, could end up as paperclips (Bostrom, 2014, p.150). The potential for uncontrollable AI systems presents novel questions of liability, duty of care, and risk allocation that existing legal frameworks are ill-equipped to address. While existential threat scenarios remain speculative, the precautionary principle would suggest that regulatory frameworks should anticipate containment challenges.

All the examined scenarios present humanity with dire scenarios, all quite plausible outcomes, once AI systems become the norm, rather than the exception in our everyday life. An argument could be made, and in fact, has been made, that such great risks, existential even, are a substantial reason for humanity to steer clear of such technological advancements, echoing the famous phrase attributed to William C. Taylor «just because you can doesn't mean you should»[7]. However, such considerations are well outside the scope of this thesis, as we believe that the creation of AI systems is an inevitable outcome under the present techno-economic complex. If so, how can these harrowing scenarios be evaded? The answer might just be Responsible AI.

3. RESPONSIBLE AI

The research and the discourse around responsibly creating AI systems have mainly focused on a few principles that seem to act as the indisputable foundation for such responsible AI systems. Large corporations, governments and international organizations seem to focus on principles such as transparency, accountability, fairness, human oversight, safety and data protection when searching for the answer to the million-dollar question of safe AI development. Thus, these pillars warrant a closer look.

[7] https://hbr.org/2011/12/just-because-you-can-doesnt-me

3.1 Transparency and Accountability in Responsible Artificial Intelligence

One of the main concerns surrounding the development of AI systems revolves around their opacity. Most AI systems use complex, multilayer, nonlinear algorithms that transform the input data in such complex and unpredictable ways that they don't allow the human engineer to understand the system's methodology. For example, Deep neural networks with their multi-layered calculations, or systems enforcing reinforcement learning methods, based on trial-and-error principle, allow for little to no understanding of the process by which they arrived at a certain output. Such opacity has led to the characterization of AI models as Black Boxes. (Felzmann, 2020 pg. 3348)

Subsequently, it is only logical that transparency has emerged as a foundational principle in discussions surrounding Responsible Artificial Intelligence, particularly about the clarity of AI-driven decision-making and the allocation of responsibility and accountability. Our present inability to understand the automated decision-making processes of AI systems poses a significant risk of diminishing transparency and therefore accountability when it comes to AI systems, simultaneously rendering human oversight largely ineffective. This challenge underscores the necessity for explainable AI (XAI), wherein the rationale behind AI-generated outcomes, as well as what data was used to inform them and how they were used, will be understandable by a human actor.

Despite its importance, explainability in AI presents notable challenges. Turning the complex mechanisms of algorithmic decision-making into an intelligible and structured format remains a complex task (Kaadoud, Fahed, Lenca. 2021). Furthermore, AI systems continuously evolve, complicating efforts to ensure consistency in explanations. Explainability must also be context-sensitive, considering factors such as the nature of the AI decision, its urgency, and the criticality of its impact. Not to mention that the effectiveness of AI explanations is contingent upon human interpretation. The explainability of AI systems isn't a problem with a single unknown factor (Crootof, Kaminski, Price II, 2023, p.434). Meaning that in addition to not knowing how to interpret the algorithms' processing mechanisms, we also don't know how individuals comprehend and assess different explanatory models. (Mikalef, 2022, p.261)

Beyond explainability, transparency in AI extends to traceability and communication, both of which are critical for legal and governance frameworks. To be able to attribute liability and correct possible AI systems' errors, we need to be able to trace AI decision-making processes. This

underscores the necessity of the establishment of AI governance practices that assign responsibility at various stages of AI deployment, ensuring that errors can be identified, assessed, and mitigated (Mikalef, 2022, p.261).

Following in the steps of Privacy by Design, Transparency by Design is gaining traction. The efforts to establish transparency shall be undertaken simultaneously with the first steps taken to create the AI system and not be addressed after its design. Furthermore, transparency measures shall account for the way AI systems combine datasets in unexpected ways, reaching unpredictable conclusions. To this end, comprehensive documentation throughout an AI system's lifecycle helps engineers better audit the systems, a key requirement for regulatory compliance and accountability. The information documented shall be able to answer questions like what data is processed, how it is processed and why a certain decision was made by the system (Felzmann, 2020 pp. 3344-3350). Navigating AI's black box won't be an easy task, and many scientists consider it an impossible one, however, it is still and shall remain in the heart of the regulators' agendas.

3.2 Accountability in Responsible Artificial Intelligence: Legal and Organizational Considerations

Accountability is another central theme in the legal and ethical discourse on responsible Artificial Intelligence, particularly in contexts where AI-driven decisions have significant consequences, such as healthcare and autonomous vehicles. A common challenge in these areas is the ambiguity of accountability, as AI systems often lack clear attributions of intent, motive, and rationality. Moreover, because AI systems are largely based upon predictions, they can easily avoid accountability (Agrawal, Gans, Goldfarb, 2022, p.60).

Accountability poses such a thorny issue as the decisions regarding it can highly alter the stakes of AI system development and deployment. The potential negative consequences of AI accountability—where responsibility may be diffused or obscured—necessitate a principled approach to AI design and deployment to ensure that accountability is both well-defined and enforceable.

Researchers have emphasized the importance of procedural mechanisms for integrating accountability into AI-based decision-making. For instance, some advocate for AI systems that

are beneficial, explicable, and transparent, urging organizations to embed accountability anchors in AI-integrated business analytics systems (Mikalef, 2022, p.262). These discussions highlight the urgent need for accountability frameworks that guide AI deployment in organizational settings.

A key question is how to distribute responsibility among the main stakeholders throughout the AI development lifecycle. It is essential to distinguish between responsibility and accountability and to analyse how these concepts influence organizational workflows and decision-making structures within both business and IT teams. Additionally, AI governance must incorporate best practice approaches to ensure that both AI systems and their development processes remain auditable and compliant with legal and ethical standards. There is no easy way to determine who will bear the burden of AI mistakes, especially when the stakes are high. A real-life manifestation of these considerations unfortunately occurred in 2016, when a fatal car crash involving an autonomously driving Tesla car occurred (Nida-Rümelin, Winter, 2024, p.5).

Finally, the question of accountability is closely linked to AI testing and safety assurance. Determining how AI systems can be assessed at various stages of development so as to ensure they meet necessary safety thresholds and do not pose risks is of great concern in AI governance. Establishing legal and regulatory guidelines for AI testing, evaluation, and compliance will be crucial in mitigating risks and ensuring that AI systems operate within ethical and legal boundaries (Mikalef, 2022, p.261).

3.3 **Human Oversight in AI Governance: Addressing Known and Unknown Risks**

A fundamental concern in responsible AI governance is determining the appropriate scope and effectiveness of human oversight in AI-driven decision-making. There are three different ways in which such human oversight manifests. Human in the loop (HITL), human on the loop (HOTL) and human in command (HIC). The HITL approach is highly invasive, granting the human the ability to interfere on every step of the way. HOTL vests the human with intervention authority during the design of the model, reducing him moving forward, to a mere monitoring authority when the system is active (AI HELG, 2019, p.16)

For the most part, these oversight mechanisms are reactive and focus on human intervention only in cases of abnormal, deviant, or unexpected AI behaviour. These approaches assume that AI-induced failures are anomalies, whereas systemic issues, such as bias or technical errors, may

drive multiple decisions without manifesting as concrete, detectable problems. Consequently, systemic issues might go unnoticed, causing harm in a subtle but accumulative fashion, instead of manifesting as a one-time incident.

Finally, the HIC mechanism grants humans the role of the general overseers. It focuses on judging the long-term effects of the system's usage and therefore, allowing the overseer to decide whether the system shall be used in a specific scenario or override certain decisions. It is generally considered as the option that better mitigates the risks presented by HITL and HOTL as it is more proactive rather than reactive (AI HELG, 2019, p.16) as it enables a dual-layered oversight approach that not only accounts for known issues—such as bias, fairness, and specific technical flaws—but also systematically monitors for unknown or unprecedented AI behaviours that may lead to unforeseen negative consequences.

The main challenge lies in structuring regulatory and organizational policies that ensure AI oversight remains dynamic, adaptive, and capable of identifying both anticipated and unforeseen risks. This requires a shift from traditional reactive corrective compliance models toward continuous evaluation mechanisms, ensuring that human oversight functions as an active safeguard rather than a mere fail-safe for AI systems. The now infamous «just pull the plug» approach when it comes to oversight mechanisms, has long become obsolete, as it requires to first be able to discern that a problem exists.

3.4 **Fairness and Bias in Responsible AI Governance**

The issues of fairness and bias have been widely discussed in the contemporary AI literature, particularly in relation to algorithmic discrimination and its legal, ethical, and societal implications. It constitutes an integral part of any Responsible AI conversation. Within this discourse, Akkerman (2024, p.72) provides empirical evidence of algorithmic bias and injustice, underscoring the challenges AI systems face in ensuring equitable outcomes, citing incidents in the Netherlands where minor administrative mistakes by the taxpayers would lead to the assumption of fraud.

The main source of discrimination in AI systems can be traced back to the datasets used to train said systems. Biased training datasets can lead to repetitively skewed decision-making in the long run for any AI system, no matter how well designed it otherwise is (Pitoura, 2020, p.4). Such

indirect biases are extremely dangerous as the cause of the problem can be intricately disguised. Ensuring that the training datasets are clear of biases is a necessary first step to achieve fair AI systems. However, it is much easier said than done as the datasets are massive and the bias might not always be as straightforward as the ethnicity of the subject or their social background.

Such indirect biases, however, are not the only factors that shall be taken into consideration when discussing AI fairness. As important, although easier to spot and therefore remedy, are direct forms of bias caused by certain centres of advanced influence. Namely, stakeholders such as companies commissioning the development of certain AI systems must abstain from integrating their self-interest in the decision-making, for the systems to remain unbiased (Mikalef, 2022, p.261).

Another factor we need to keep in mind when characterising an AI system as fair or not, is its accessibility. AI systems need to be designed for the greater societal good and not for the enrichment of a certain country, company, or any other singular entity. Consequently, they shall be designed in a way that grants access to the widest amount of people possible.

These insights suggest that AI fairness assessments should extend beyond isolated decision-making instances and consider the long-term impact of bias over an individual's life course. Additionally, regulators and policymakers must mandate fairness testing across multiple points of a system's lifetime to better assess AI's long-term implications on social and economic equity (Leslie, 2023, p.12).

3.5 AI Design and Security in Responsible AI Governance

In light of the growing ubiquity of AI, it is imperative that AI applications are designed with an acute awareness of ethical principles and security measures. As previously noted, ethical dilemmas are integral to the development of AI systems, as these systems must navigate decisions that reflect the moral and ethical standards upheld by the societies in which they operate. In addition to addressing ethical concerns, there are critical considerations related to the security of AI systems.

As AI applications become increasingly integrated into high-stakes domains, such as healthcare, finance, and autonomous transportation, it is essential to establish robust security policies that

safeguard the integrity of AI systems. These policies must span multiple stages of the AI lifecycle, from initial design to deployment and post-deployment monitoring.

When considering the security of AI systems, one needs to keep in mind the additional threats that other AI systems can pose if used in an adversarial way, in addition to the already common threats posed by human and other artificial agents. AI systems must be tested extensively from the very early stages of their development, as to ascertain their integrity, regarding both hardware and software.

Preventive strategies and fail safes for threats such as data poisoning, model tampering, or unauthorized data access must be a key component of AI security policies. It is equally important to ensure that AI models are protected from leaks, while still maintaining the capacity to demonstrate transparency. From a legal and regulatory perspective, the development of security policies in AI must balance proactive security measures with the need for operational transparency, ensuring that AI systems can be effectively audited, and their decisions can be verified without compromising data privacy or intellectual property.

3.6 Data Governance and Privacy in Responsible AI Development

The principle of data governance is of utmost importance for the responsible development and deployment of AI systems. It is found at the very core of any AI system as it must be present from the creation and collection of data to the usage of said data to train and reinforce AI systems, to its eventual use in AI automated decision-making. The case study of the Australian government's "Robodebt" program, designed to automatically calculate and collect welfare overpayment debts, resulted in severe distress for citizens and welfare agency staff due to errors in data handling and algorithmic miscalculations (Mikalef, 2022, p.262).

Similarly, the case study of a data analytics project in Denmark's healthcare sector gives us valuable insights from its emergence, expansion, and eventual collapse. The study underscores the need for strong governance frameworks that can manage the dynamics between functional, stakeholder, and data subsystems to ensure a balanced approach that supports long-term project success.

Beyond the case studies referenced, data governance in AI also involves addressing critical issues related to privacy and data protection. These include safeguarding data against unauthorized

access, ensuring data integrity, and maintaining its high quality. One key challenge remains understanding how to effectively manage socially constructed biases, inaccuracies, errors, and mistakes in data, and how to ensure that AI systems operate based on accurate and representative data. At the same time, organizations must strike a delicate balance, as over-cleaning data could result in the sacrifice of quality outcomes.

Another pressing issue is the biased tagging of data used in AI training. Early research has highlighted the risks associated with biased data labelling, which can significantly influence AI outcomes. Other data governance issues, already familiar but greatly amplified due to the scale of data used by AI systems and the innovative ways in which they are combined and used, have to do with the who has access to the data, how the subject can know which of their data are being used and how these data can be deleted.

In summary, a comprehensive data governance framework is critical to ensure that AI technologies are designed and deployed in ways that prioritize ethics, privacy, and fairness while maintaining high data quality and security. As AI continues to evolve, ongoing research and dialogue will be needed to refine these governance practices and address emerging challenges, particularly concerning bias and risk perception.

4. REGULATORY FRAMEWORK

The abovementioned constitute some of the commonly accepted fundamental and indisputable principles that shall guide the creation of responsible AI models. In the paragraphs that follow, we will examine how regulators have sought to enforce said principles by integrating them in their Legal frameworks, both enforced and proposed, and in their technical guidelines.

4.1 The Legal Basis and Scope of the Artificial Intelligence Act (AIA)

The Regulation (EU) 2024/1689 of the European Parliament and of the council of 13 June 2024, commonly referred to as «The Artificial Intelligence Act (AIA)» has its legal grounds in Article 114 of the Treaty on the Functioning of the European Union (TFEU)[8]. This article is part of the EU's broader single market strategy in the digital age and the New Legislative Framework

[8] AI Act, third remark

(European Commission, 2023), which aims to create harmonized rules essential for the proper functioning of the internal market. The primary goal of these harmonized rules is to avoid fragmentation within the EU market and eliminate the legal uncertainty stemming from the current patchwork of national laws (European Commission, 2021, p. 8). According to the European Commission, the AIA ensures a level playing field, better safeguards people's rights, and enhances Europe's competitiveness and innovation, particularly in AI systems. The regulation also plays a critical role in protecting the EU's digital sovereignty and reinforcing its regulatory powers to shape global standards (European Commission, 2021, p. 11).

The development of the AIA follows extensive stakeholder consultation, incorporating feedback from the High-Level Expert Group on AI and building on the EU White Paper[9]. The AIA complements other important regulations, including the General Data Protection Regulation (GDPR), the Digital Services Act (DSA), the Digital Markets Act (DMA) and various other regulatory frameworks such as the Data Governance Act, the AI Liability Directive, and the Revised Product Liability.

The AIA is a comprehensive regulation, consisting of multiple chapters, articles and annexes. Chapter I outlines the subject matter and scope of the Act, while providing useful definitions. According to Article 1, the AIA lays down harmonized rules for the placing on the market, putting into service, and use of AI systems within the Union. This includes, prohibiting certain AI practices, placing specific requirements for high-risk AI systems and the obligations of operators, transparency rules for AI systems intended to interact with natural persons, such as emotion recognition and biometric categorization systems, rules for AI systems that generate or manipulate image, audio, or video content, market monitoring and surveillance requirements (Wörsdörfer, 2023, p.109).

As per Article 2, the AIA applies to both providers of AI systems, deployers, importers and distributors, whether established or located within the EU or in a third country, as long as the AI system is put into service, placed in the market or its output is used in the European Union. A couple of interesting exemptions from the scope of the Regulation include AI systems used exclusively for military, defence or national security purposes, echoing the Union's desire to stay clear of regulatory red tape when it comes to its sovereignty. Secondly, the Regulation, for the

[9] White Paper on Artificial Intelligence: a European approach to excellence and trust - European Commission

most part, does not apply to AI systems released under free and open-source licenses, underlying the EU's commitment to promote the creation of open-source systems.

Furthermore, the regulation defines AI as «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» (Art. 3.1). In summary, the AIA represents a significant step toward regulating AI within the EU, ensuring that AI technologies align with ethical principles, protect citizens' rights, and foster innovation while maintaining legal clarity and consistency across the internal market.

4.2 The Artificial Intelligence Act and the implementation of the main principles for responsible AI

The Artificial Intelligence Act (AIA) operates within a precautionary, risk-based framework, as detailed in Chapters II, III, and IV of the Regulation. The AIA categorizes AI systems into three distinct risk levels—unacceptable risk, high risk, and low or minimal risk—which are depicted in what is referred to as the "pyramid of risks." The regulation takes a strong stance on unacceptable risk AI systems, prohibiting their market access altogether due to their inherently high risk that cannot be mitigated by risk-reducing interventions. However, some have commented that the list of prohibited systems might actually be too short (van der Sloot, 2024, p.154)

For high-risk AI systems, the AIA permits market access, but only under strict compliance with its provisions. This includes meeting ex-ante technical requirements (i.e., requirements to be fulfilled before the system is put on the market) and undergoing ex-post market monitoring to ensure ongoing compliance. On the other hand, low-risk or minimal-risk AI systems are subject to more general safety standards.

In essence, the AIA adheres to the principle of proportionality, which means that it places the majority of its regulatory focus on high-risk AI systems, imposing stringent requirements to ensure their safety and compliance (Wörsdörfer, 2023, p.109). For non-high-risk systems, the regulation adopts a more lenient approach, imposing limited transparency obligations, such as voluntary codes of conduct, to avoid imposing undue burdens, particularly on small and medium-sized enterprises (SMEs). The goal is to balance the need for regulation with the desire

to minimize compliance costs and avoid stifling innovation within smaller organizations (European Commission, 2021a, p. 7).

Article 5 of the Artificial Intelligence Act (AIA) provides a list of prohibited AI systems that pose an unacceptable risk or violate EU values, particularly those that infringe on fundamental rights. The Regulation prohibits manipulative, deceptive systems aiming to distort the behaviour of persons and systems that take advantage of subliminal techniques, altering people's decision-making, leading to the cause of harm. Moreover, the AIA prohibits systems that exploit vulnerable groups (e.g., children, the elderly, people of a specific socio-economic background or individuals with physical or mental disabilities), altering their behaviour and causing psychological or physical harm.

Further, Article 5 specifically bans AI-based social scoring leading to detrimental or unfavourable treatment of certain persons. This includes practices similar to China's Social Credit System, which generates trustworthiness or citizen scores. Additionally, the use of real-time remote biometric identification systems (such as facial recognition software in large-scale CCTV networks) for law enforcement purposes in publicly accessible spaces is also prohibited, as it raises significant concerns about privacy, freedom, and surveillance, unless certain, limited objectives deem such use strictly necessary. These are some of the prohibited systems, an exclusive list of which is listed on Article 5 of AIA.

The AIA provides certain exceptions to these prohibitions. For example, law enforcement agencies may still use these technologies under specific circumstances, such as for searches related to serious criminal offenses (Article 5 of AIA), but these exceptions are tightly regulated and must adhere to strict conditions to safeguard fundamental rights. These exceptions allow for the use of AI technologies, such as real-time biometric identification, in cases where there is a threat to life or physical safety of a person, such as when detecting, localizing, identifying, or prosecuting perpetrators or crime suspects. However, the use of such AI systems for law enforcement purposes requires authorization by either a judicial or an independent administrative authority to ensure accountability and prevent misuse.

Additionally, we shall keep in mind that, as stated before, the AIA specifies that AI systems developed or used exclusively for military purposes are excluded from the scope of the regulation. This exclusion ensures that the AIA does not interfere with national defence or military operations, even when we are dealing with otherwise prohibited systems.

Chapter III of the Artificial Intelligence Act (AIA) addresses high-risk AI systems—those that may pose significant risks to individual health and safety or infringe upon fundamental rights. According to Article 6 and Annex III, high-risk AI systems include, but are not limited to, AI systems used for remote biometric identification, emotion recognition, as safety components for critical digital infrastructure, road traffic, or in the supply of essential utilities, such as water and electricity. Moreover, AI systems are considered high-risk when they are used to determine access or admission or to assign natural persons to educational and vocational training institutions at all levels, to evaluate learning outcomes, to recruit persons, to filter job applications[10], to evaluate candidates, to determine promotions, to establish credit score and many more systems. Art. 6 also makes provisions in its third paragraph to exclude systems for the high-risk category, despite being referred to in Annex III. Expectedly, this list is subject to updates, with additional AI systems potentially classified as high-risk (Art. 7 AIA), as the European legislator can't possibly foresee every type of high-risk system that will be designed.

Once a system is categorized as high-risk, its providers must adhere to various mandatory requirements under Section 2. In those requirements, we start to see how the European legislator has opted to promote the main principles for creating responsible AI systems. These include the establishment of a risk management system and impact assessments (Art. 9), data governance and management (Art. 10), and the preparation of technical documentation (Art. 11). Providers must also ensure traceability and record-keeping (Art. 12), provide transparency and user information (Art. 13), facilitate human oversight (Art. 14), and ensure accuracy, robustness, and cybersecurity (Art. 15). To further understand how the European legislator wishes to adhere to the principles of responsible AI we shall expand upon some of the charted requirements.

4.2.1 Requirements for high-risk AI systems

As previously discussed, data governance is paramount to establishing responsible AI systems. Subsequently, the AIA lays out a comprehensive data governance framework. Art. 10 requires that data sets used for the various stages of an AI system's development (i.e. training, validation and testing) shall meet certain quality criteria. In order for the desired data quality to be achieved, certain practices need to be adopted to make sure that, among other things, the data origin is

[10] Amazon's algorithm in 2015 discriminated against women by excluding them from technical job applications. (Nida-Rümelin, Winter, 2024, p.5).

known, the data sets are labelled, clean, updated, the intended data representations are clear and they have been examined for possible biases that could result to unfair AI systems. Moreover, extensive provisions are made for the creation of bias detection and prevention mechanisms, while also trying to make sure that the subjects of personal data are sufficiently protected.

Articles 11, 12 and 13 of the Regulation aim to achieve sufficient transparency levels for high-risk systems by requiring, firstly, that, before they are placed on the market or put into service, their technical documentation needs to be drawn up. The technical documentation must provide sufficient information to demonstrate that the system complies with all the requirements set forth in the relevant articles of the Regulation. Secondly, that high-risk AI systems must incorporate automated log keeping over their lifetime. Such logs shall include events important to assessing risks and monitoring the operation of the system. Finally, requires that high-risk AI systems are designed in a sufficiently transparent way, meaning that the system's output shall be explainable. Additionally, art. 13 deems necessary to accompany high-risk AI systems with detailed instructions, including any foreseeable circumstance that can lead to the manifestation of risks.

Art. 14 introduces the human in the equation. Art. 14 does not determine how the human shall be able to interfere with the system but rather lays out the requirements for said human oversight to even be possible in the first place. It stipulates that human-machine interface tools shall be included in the design of high-risk systems to enable oversight. Such tools shall allow for the interpretation, reversal, overriding of any given output and a failsafe for the interruption of the system's operation.

At last, Art. 15 attempts to quell accuracy, robustness and cybersecurity concerns. The systems must perform at least in par with predetermined baselines for the entirety of the life cycle. The systems shall be protected from data and model poisoning attacks, adversarial examples and model evasion, confidentiality attacks and model flaws.

Additionally, the providers of high-risk AI systems are required to, affix the CE marking to indicate compliance with the AIA (Art. 16, 49), set up a quality management system (Art. 17), conduct and maintain relevant conformity assessments (Art. 19), automatically generate AI system logs (Art. 19), take corrective actions if non-compliance is identified (Art. 20), inform national authorities about any significant changes, risks, or corrective actions, cooperate with competent authorities by providing access to necessary documentation and system logs (Art. 21)

The conformity assessment procedure is designed to ensure high-risk AI systems comply with AIA standards. Notably, these systems must undergo a new assessment if they are substantially modified (Art. 43). Once passed, the systems receive a certificate of compliance (Art. 44), valid for up to five years, four if the system is high-risk, with the possibility of renewal, suspension, or withdrawal (Art. 45).

Before high-risk AI systems are introduced to the market, providers must also register them in a newly established E.U.-wide database. This database, managed by the European Commission, ensures transparency by making the information publicly accessible (Art. 49).

Title IV of the Artificial Intelligence Act (AIA) addresses low-risk and minimal-risk AI systems, which, while not posing significant risks, are still subject to specific transparency obligations to ensure responsible use. Such low or minimal-risk AI systems include systems interacting with humans, such as LLMs like Chat GPT, DeepSeek and the other famous AI systems of today. Although the application of such systems hasn't been without its complaints so far (CAIDP, 2023, OPEN AI, FTC 2023, Complaint). For these systems, it is mandatory to ensure that natural persons interacting with them are informed of the nature of the interaction. Specifically, providers must include clear disclosures that users are interacting with an AI system (e.g., bot interaction), the system involves emotion recognition or biometric categorization, the system is generating or manipulating synthetic content or deepfake media.

These transparency requirements aim to ensure that users are aware of their interaction with AI and to prevent misuse or unintentional manipulation, especially when it comes to sensitive technologies like emotion detection and media manipulation.

4.2.2 Regulatory Sandboxes

One key aspect of the Artificial Intelligence Act (AIA) is its commitment to fostering AI innovation while ensuring regulatory oversight. To achieve this, AI regulatory sandboxes (Art. 57) are introduced as a mechanism to support the development and testing of AI technologies in a controlled environment, ensuring compliance with the AIA before these systems are marketed.

Art 57 stipulates that each Member State must establish at least one operational AI regulatory sandbox by August 2, 2026, allocating sufficient resources for their implementation, while also providing exemptions in cases where every member doesn't need to have one, if they can be

hosted in another existing sandbox effectively. Testing, within the sandboxes, can provide developers of AI systems with the chance to face real–world conditions while being supervised, before placing their systems in the market. Supervisory authorities and governing bodies will be providing guidance and support to the developers to optimise each AI system. Following the successful participation of a system developer in a sandbox, the competent authorities shall provide relevant written proof and an exit report. Both can be used to attest to the compliance of the tested system with the Regulation. Spain's government, in partnership with the European Commission, launched the EU's first AI regulatory sandbox pilot in Brussels back in October 2022[11].

4.2.3 Governance Structure of the EU AI Act: Key Features

The EU AI Act establishes a comprehensive governance framework combining Union-level and national authorities. The AI Office (Article 64) serves as the Commission's hub for developing Union expertise and capabilities in AI, with Member States facilitating its tasks.

The European Artificial Intelligence Board (Articles 65-66) brings together representatives from all Member States to ensure consistent application of the regulation. The Board advises on implementation, coordinates national authorities, shares expertise, develops common understandings, and contributes to enforcement standards —particularly for general-purpose AI models. It establishes standing sub-groups for market surveillance and notified bodies. The EAIB aims to ensure coherence in regulatory oversight and will likely become a key body for overseeing AI regulations within the EU.

An Advisory Forum (Article 67) provides technical expertise to both the Board and Commission. Its membership balances industry (including startups and SMEs), civil society, and academia, with permanent representation from standardization bodies and EU agencies. The forum prepares opinions and recommendations on request.

A Scientific Panel of Independent Experts (Article 68-69) supports enforcement activities with objective technical expertise, particularly regarding general-purpose AI models. The panel alerts authorities about potential systemic risks, assists with technical evaluations, and supports cross-border surveillance. Member States can access this expert pool for their enforcement activities.

[11] https://digital-strategy.ec.europa.eu/en/news/first-regulatory-sandbox-artificial-intelligence-presented

This multi-layered structure ensures balanced stakeholder input, technical expertise, and coordination between national and Union-level implementation of the AI Act.

4.2.4 Post-Market Monitoring

The AIA includes post-market monitoring mechanisms to track the performance of AI systems even after they are marketed. Under Chapter IX, AI system providers must establish a post-market monitoring system (Art. 72) that allows them to collect and analyse data regarding system performance and compliance with the AIA. Serious incidents, system malfunctions, or breaches of compliance must be investigated and reported to relevant authorities.

National supervisory authorities are responsible for market surveillance, ensuring AI systems remain compliant throughout their lifecycle. These authorities can also request access to system logs or, in certain cases, the source code of AI systems (Art. 75). If a system fails to comply with AIA standards, authorities may require the provider to withdraw the system from the market or recall it (Art. 79).

4.2.5 Administrative Fines

Under Chapter XII, the AIA includes provisions for penalties against non-compliance. These fines are designed to be effective, proportionate, and dissuasive to ensure stakeholders respect the rules. Penalties for violations can reach up to EUR 35 million or 7% of the total worldwide annual turnover, depending on the severity of the infringement (Art. 99).

4.3 **The other side of the Atlantic**

In the U.S., AI legislation has evolved with contrasting approaches across different administrations. Unlike the European Union, which has put into effect a comprehensive AI regulatory framework, the U.S. does not have a singular, overarching AI Act. Instead, its strategy consists of fragmented policies aimed at promoting innovation while trying to manage the associated risks. Below are some key developments in federal AI legislation.

4.3.1 *The National Artificial Intelligence Initiative Act (NAII) of 2020*

Signed during President Trump's first term, the NAII aims to drive AI innovation in key sectors. It establishes a national strategy for AI development, emphasizing investment in AI research and education. The act also promotes collaboration between federal agencies, academic institutions, and the private sector to advance AI technologies while ensuring the U.S. remains a leader in global AI innovation[12].

4.3.2 *The Bipartisan House Task Force Report on AI (2024)*

Published in December 2024, this bipartisan report presents guiding principles, key findings, and recommendations for U.S. policymakers on AI. It is a forward-looking document that outlines the necessary steps Congress should take to keep pace with advancements in AI technology, balancing innovation with ethical concerns and risks. It aims to help shape future legislative efforts on AI regulation[13].

4.3.3 *President Joe Biden's Executive Orders (2023-2025)*

Former President Biden made significant moves to regulate AI risks while encouraging ethical and responsible use. The Executive Order on Safe, Secure, and Trustworthy AI (2023) focused on establishing principles for the safe and secure development of AI systems, emphasizing privacy, fairness, and accountability[14].

Furthermore, Biden's administration introduced the AI Bill of Rights as a framework for ensuring that AI technologies respect civil liberties and fundamental rights[15].

However, many of these policies were revoked when President Trump took office in January 2025. Still, some initiatives, such as Executive Order 14141 (on AI infrastructure) and Executive Order 14144 (on cybersecurity), were maintained in some form under Trump's administration.

[12] https://www.congress.gov/bill/116th-congress/house-bill/6216
[13] https://science.house.gov/2024/12/house-bipartisan-task-force-on-artificial-intelligence-delivers-report
[14] https://www.federalregister.gov/documents/2023/11/01/2023-24283/safe-secure-and-trustworthy-development-and-use-of-artificial-intelligence
[15] https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights/

4.3.4 *State-Level AI Legislation*

In addition to federal efforts, state governments in the U.S. are actively shaping AI regulation through state-level legislation. Given the decentralized nature of U.S. governance, much of the actionable AI legislation is emerging at the state level, where jurisdictions are crafting laws to manage the impact of AI technologies.

The Colorado AI Act of 2024, inspired by the EU AI Act, uses a risk-based approach to regulate high-risk AI systems. This act emphasizes the principle of transparency stating that AI system operators must provide clear information about how their systems work and the risks involved. Moreover, it requires that companies maintain comprehensive documentation of their AI system development processes and AI providers to implement measures to identify and reduce risks associated with their technologies (Jariwala, 2024, p.25). By drawing from international best practices like the EU AI Act, Colorado aims to set a high standard for AI governance at the state level.

The Illinois Supreme Court's AI Policy, introduced on January 1, 2025, focuses on integrating AI into the judicial system responsibly. The key goals of this policy include ensuring accountability and ethical use of AI in judicial proceedings, protecting judicial integrity by providing guidelines on how AI systems can assist in but not replace human decision-making in legal processes[16]. This move positions Illinois as a leader in judicial AI governance, emphasizing the importance of using AI in ways that respect legal rights and human judgment.

In the U.S., AI legislation remains fragmented, with significant federal initiatives like the National AI Initiative Act and various Executive Orders under Presidents Biden and Trump. At the same time, states like Colorado and Illinois are stepping in to create their own regulations, addressing areas where federal laws may be lacking. These state-level efforts, especially those inspired by the EU AI Act, offer valuable insights into how AI regulation might evolve in the U.S. over time.

For business leaders and AI developers, understanding both federal and state regulations is essential for ensuring compliance and mitigating the risks associated with AI deployment, particularly as state-level laws could set a precedent for future national policies.

[16] https://www.illinoiscourts.gov/News/1485/Illinois-Supreme-Court-Announces-Policy-on-Artificial-Intelligence/news-detail/

However, such calculations could be proven incorrect as the political climate in the USA might shift following the re-election of President Donald J. Trump. The deep-rooted fear of American politicians when it comes to Chinese technologies and the national security of the United States of America, previously witnessed in the social media arena, has already shifted in the theatre of AI development.

4.3.5 Decoupling America's Artificial Intelligence Capabilities from China Act of 2025

The proposed bill, titled the "Decoupling America's Artificial Intelligence Capabilities from China Act of 2025," seeks to amend Title 18 of the United States Code to restrict U.S. persons from advancing artificial intelligence (AI) and generative AI capabilities within the People's Republic of China (PRC). The bill introduces several key provisions aimed at curbing the transfer of AI-related technology, intellectual property, and research to Chinese entities, particularly those associated with the PRC's military-civil fusion strategy, which integrates civilian resources for military purposes[17].

Key provisions include prohibitions on Import and Export. The bill prohibits the importation of AI or generative AI technology and intellectual property developed or produced in the PRC into the U.S., and similarly bans the export, reexport, or in-country transfer of such technology to or within the PRC. Violations of these prohibitions are subject to both criminal and civil penalties, including fines and forfeiture of federal benefits.

Additionally, it sets forth restrictions on research and development, stating that U.S. persons are barred from conducting, transferring, or collaborating on AI or generative AI research and development with Chinese entities of concern, which include educational institutions, research labs, corporations, and government entities affiliated with the PRC. Penalties for violations include substantial fines, forfeiture of federal contracts or grants, and ineligibility for federal financial assistance for up to five years.

The prohibitions don't stop here. The bill also prohibits U.S. persons from holding or managing financial interests in, or providing financing to, Chinese entities involved in AI research and development that support the PRC's military-civil fusion strategy, surveillance capabilities, or

[17] S.321 - 119th Congress (2025-2026): Decoupling America's Artificial Intelligence Capabilities from China Act of 2025 | Congress.gov | Library of Congress

human rights abuses. Violations are subject to penalties under the International Emergency Economic Powers Act.

The bill mandates the issuance of regulations by the Secretary of Commerce and the Attorney General, in consultation with other federal agencies, to enforce these prohibitions and ensure compliance.

In summary, the bill represents a significant legislative effort to decouple U.S. AI capabilities from China, particularly in areas that could enhance the PRC's military or surveillance capabilities. It reflects growing concerns over the strategic competition between the U.S. and China in the realm of advanced technologies and aims to safeguard U.S. technological superiority and national security interests.

4.3.6 «Removing barriers to American leadership in artificial intelligence»

This executive order (Executive Order 14179), issued by the President, aims to reinforce the United States' global leadership in artificial intelligence (AI) by removing barriers to innovation and ensuring that AI development remains free from ideological bias or engineered social agendas. The order reflects a policy shift to prioritize American dominance in AI, emphasizing its role in promoting human flourishing, economic competitiveness, and national security.

The order underscores the U.S.'s historical leadership in AI innovation, driven by free markets, research institutions, and entrepreneurship. It seeks to maintain this leadership by revoking existing AI policies that hinder innovation and by promoting the development of unbiased AI systems.[18].

The policy of the U.S. is to sustain and enhance its global AI dominance to benefit human flourishing, economic competitiveness, and national security. The Order sets forth to develop an AI Action Plan. Within 180 days, key White House advisors, including the Assistant to the President for Science and Technology (APST), the Special Advisor for AI and Crypto, and the Assistant to the President for National Security Affairs (APNSA), are tasked with developing and

[18] S.321 - 119th Congress (2025-2026): Decoupling America's Artificial Intelligence Capabilities from China Act of 2025 | Congress.gov | Library of Congress

submitting an action plan to achieve the policy goals outlined in the order. This plan will be developed in coordination with other relevant agencies and advisors.

Simultaneously, it revokes previous AI policies, mandating an immediate review of all policies, directives, regulations, and actions taken under the revoked Executive Order 14110 (Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence). Any actions inconsistent with the new policy must be suspended, revised, or rescinded. Agencies are instructed to provide exemptions where necessary until final actions can be taken.

This executive order best represents the significant shift in U.S. AI policy, prioritizing innovation and global leadership over previous regulatory frameworks that may have been perceived as overly restrictive. By revoking Executive Order 14110 for Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence[19], the administration aims to create a more permissive environment for AI development, free from constraints. The focus on removing barriers to innovation aligns with broader goals of economic competitiveness and national security, reflecting a strategic response to global technological competition, particularly with China. To quote the sitting President "It is the policy of the United States to sustain and enhance America's global dominance in order to promote human flourishing, economic competitiveness, and national security"[20].

However, the order's emphasis on eliminating "ideological bias" and "engineered social agendas" raises questions about the potential deprioritization of ethical considerations in AI development, which were central to the revoked Executive Order 14110. The long-term implications of this policy shift on AI governance, ethical standards, and international collaboration remain to be seen.

4.4 Technical Guidelines

Executive Orders, Regulations, national and international legislations are not the only AI related regulatory initiatives to date. Maybe even more important, especially in the early stages of

[19] Federal Register :: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence
[20] S.321 - 119th Congress (2025-2026): Decoupling America's Artificial Intelligence Capabilities from China Act of 2025 | Congress.gov | Library of Congress

legislative regulation, technical standards and guidelines have been issued by esteemed organizations to help AI developers and providers adhere to the Responsible AI pillars.

4.4.1 ISO/IEC 23894:2023 standard

The ISO/IEC 23894:2023 standard, titled Information Technology — Artificial Intelligence — Guidance on Risk Management, provides a comprehensive framework for managing risks associated with the development, deployment, and use of artificial intelligence (AI) systems. This standard is particularly relevant to the discourse on AI safety and responsible AI, as it emphasizes the integration of risk management principles into AI-related activities to mitigate potential harms and ensure ethical, transparent, and accountable AI practices[21].

The standard builds on the foundational principles outlined in ISO 31000:2018, which emphasizes an integrated, structured, and comprehensive approach to risk management. For AI systems, these principles are adapted to address the unique challenges posed by AI, such as transparency, explainability, and increased risk possibilities.

A key principle is the inclusivity of stakeholders, which is critical for AI systems given their potential far-reaching impacts. The standard highlights the importance of engaging diverse stakeholders—including regulators, business owners, and end-users—to identify risks related to data collection and processing, algorithmic bias, and fairness. This aligns with the broader ethical imperative of ensuring that AI systems do not perpetuate or exacerbate societal inequalities.

The dynamic nature of AI systems is another focal point. AI systems, particularly those involving machine learning, are inherently adaptive and continuously evolve, necessitating ongoing monitoring and risk assessment. This is crucial for maintaining AI safety, as static risk management frameworks may fail to capture emergent risks. It provides a framework for integrating risk management into organizational activities, emphasizing the role of leadership and commitment in fostering a culture of responsible AI. Top management is encouraged to communicate policies on AI risk management to stakeholders, thereby building trust and accountability.

[21] ISO/IEC 23894:2023 - AI — Guidance on risk management

The framework also underscores the need for specialized resources to manage AI risks, reflecting the technical complexity of AI systems and the expertise required to address issues such as algorithmic bias, data privacy, and security vulnerabilities. Moreover, it outlines a systematic process for managing AI risks, including risk identification, assessment, treatment, and monitoring. These processes are tailored to the AI context, with particular attention to the lifecycle of AI systems—from design and development to deployment and decommissioning.

A notable aspect is the emphasis on transparency and explainability, which are critical for ensuring that AI systems are understandable to stakeholders and can be audited for compliance with ethical and legal standards. This is particularly important in high-stakes applications, such as healthcare or criminal justice, where AI decisions can have profound societal impacts (Gillis, Laux, Mittelstadt, 2021, p.7).

The standard highlights the importance of adhering to legal and regulatory requirements specific to AI, as well as guidelines on the ethical use of AI issued by governments, standardization bodies, and industry associations. This reflects the growing recognition of the need for a harmonized approach to AI governance, balancing innovation with societal values. It also addresses the contractual implications of AI systems, particularly those involving continuous learning, which may affect an organization's ability to meet contractual obligations. This underscores the need for careful consideration of legal and ethical risks throughout the AI lifecycle.

The ISO/IEC 23894:2023 standard standard advocates for a culture of continual improvement in AI risk management, driven by ongoing learning and adaptation. This is particularly relevant in the context of AI safety, as the rapid evolution of AI technologies necessitates proactive risk management strategies that can respond to emerging threats and opportunities. It describes a robust foundation for advancing AI safety and responsible AI by embedding risk management principles into the design, development, and deployment of AI systems. By emphasizing stakeholder inclusivity, transparency, and dynamic risk assessment, the standard addresses key ethical and technical challenges associated with AI, such as bias, opacity, and adaptability and is at large, echoing the AIA positions on AI safety. Furthermore, its focus on leadership commitment and continual improvement aligns with the broader goal of fostering a culture of accountability and trust in AI systems.

4.4.2 *The ISO/IEC 42001:2023 standard*

The ISO/IEC 42001:2023 standard, titled Information Technology — Artificial Intelligence — Management System, establishes a comprehensive framework for organizations to develop, implement, and maintain an Artificial Intelligence (AI) Management System (AIMS). This standard is particularly relevant to the discourse on AI safety and responsible AI, as it provides a structured approach to managing the unique risks and ethical challenges associated with AI systems[22]. By embedding principles of accountability, transparency, and risk management into organizational processes, the standard aims to ensure that AI systems are developed and deployed in a manner that prioritizes human well-being, equity, and safety.

It is designed for organizations of all sizes and sectors that develop, provide, or use AI systems. It emphasizes the need for organizations to integrate AI management into their broader operational and governance frameworks, ensuring that AI systems align with organizational objectives and societal expectations. The standard is particularly focused on addressing the unique challenges posed by AI, such as automated decision-making, lack of transparency, and continuous learning capabilities, which necessitate specialized management practices beyond those used for traditional IT systems.

The standard underscores the importance of transparency, accountability, and ethical considerations in AI development and deployment. It requires organizations to establish clear policies and objectives for AI systems, ensuring that they are aligned with societal values and legal requirements.

The standard provides a detailed framework for AI risk assessment and risk treatment, requiring organizations to systematically identify, evaluate, and mitigate risks associated with AI systems. This includes risks related to data quality, algorithmic bias, security vulnerabilities, and societal impacts. A notable feature is the requirement for AI system impact assessments, which mandate organizations to formally evaluate the potential impacts of AI systems on individuals, groups, and society. This process is critical for ensuring that AI systems are designed and deployed in a manner that minimizes harm and maximizes societal benefit.

Furthermore, it places a strong emphasis on leadership commitment to responsible AI practices. Top management is required to demonstrate accountability by establishing clear AI policies,

[22] ISO/IEC 42001:2023 - AI management systems

allocating resources for AI risk management, and ensuring that AI systems are aligned with organizational values and objectives. The standard also highlights the role of governing bodies in overseeing AI systems, particularly in ensuring that AI-related risks are managed effectively and that the organization complies with legal and ethical standards.

The standard also emphasizes the importance of documented information and performance measurement, enabling organizations to track the effectiveness of their AI management systems and demonstrate accountability to stakeholders. It is designed to be compatible with other management system standards, such as those related to quality, safety, security, and privacy. This facilitates the integration of AI management into existing organizational frameworks, ensuring a holistic approach to risk management and governance.

To sum up, the ISO/IEC standards examined above represent a significant step forward in the global effort to ensure that AI technologies are developed and deployed in a manner that prioritizes safety, transparency, and ethical responsibility. By providing a structured framework for AI risk management, the standards address key challenges such as algorithmic bias, data quality, and societal impacts, which are critical for ensuring the responsible use of AI.

The ISO/IEC standards provide a robust foundation for advancing AI safety and responsible AI by embedding risk management and ethical principles into the design, development, and deployment of AI systems. As AI continues to permeate various sectors, the principles and processes outlined in these standards will be instrumental in guiding the responsible use of AI and mitigating potential harms, thereby contributing to a safer AI environment. The ISO/IEC standards establish a comprehensive framework for organizations to develop, implement, and maintain an Artificial Intelligence (AI) Management System (AIMS). These standards are particularly relevant to the discourse on AI safety and responsible AI, as they provide a structured approach to managing the unique risks and ethical challenges associated with AI systems. By embedding principles of accountability, transparency, and risk management into organizational processes, the standards aim to ensure that AI systems are developed and deployed in a manner that prioritizes human well-being, equity, and safety.

4.4.3 AI RMF Playbook

The AI RMF Playbook is a comprehensive guide developed by the National Institute of Standards and Technology (NIST) under the U.S. Department of Commerce. It provides a framework for managing risks associated with Artificial Intelligence (AI) systems. The playbook is structured around four core functions: Govern, Map, Measure, and Manage, which are designed to help organizations identify, assess, and mitigate risks related to AI technologies.[23].

The Playbook focuses on cultivating a culture of risk management within the organization. It includes policies, processes, and procedures for managing AI risks. Covers legal and regulatory requirements, trustworthy AI characteristics, risk tolerance, and documentation practices. Emphasizes the importance of training, executive leadership accountability, and diversity in decision-making teams.

It aims to recognize the context in which AI systems operate and identify related risks, to understand the intended purpose of AI systems, potential beneficial uses, and context-specific laws and norms. Like the ISO standards, it encourages interdisciplinary collaboration and stakeholder engagement to map risks effectively. In Addition, it documents system requirements, human-AI configurations, and potential impacts on individuals, communities, and society.

In order to achieve its goals, the Playbook focuses on assessing, analysing, and tracking identified risks, by selecting appropriate metrics for measuring AI risks, evaluating system performance, and ensuring the validity and reliability of AI systems. It emphasizes the importance of testing, evaluation, validation, and verification (TEVV) processes, while encouraging regular monitoring of AI systems in production to detect and respond to risks such as model drift or unexpected behaviour.

To manage risks, it prioritizes and acts upon risks based on their projected impact. It determines whether an AI system should proceed with development or deployment, prioritizing risk treatment, and documenting residual risks and encourages the use of non-AI alternatives when appropriate and ensures that resources are allocated effectively to manage high-priority risks. When it comes to deployed AI systems it focuses on responding to incidents, and decommissioning systems that no longer meet risk tolerance levels.

[23] NIST AI RMF Playbook | NIST

The playbook emphasizes the importance of documenting processes, risks, and outcomes to ensure accountability and transparency. Each section of the playbook provides actionable steps for organizations to implement the framework effectively. To this end, it includes references to external resources, frameworks, and best practices to support organizations in their AI risk management efforts. It is designed to be flexible and can be applied across various industries and sectors. high-stakes environments, such as healthcare, finance, transportation, and security. It emphasizes the importance of integrating trustworthy AI characteristics into organizational policies and practices.

In Conclusion, the AI RMF Playbook serves as a practical guide for organizations to manage AI risks effectively. By following the framework outlined in the playbook, organizations can ensure that their AI systems are trustworthy, transparent, and aligned with societal values and legal requirements. One of the frontrunners in AI RnD, Microsoft has publicly stated that follows the playbook (Microsoft 2023, P.11).

5. STATUS QUO AND REGULATORY IMPACT ASSESSMENT

Between voting into effect extensive Regulations, issuing grandeur Executive Orders and putting together lengthy Handbooks and Standards, it would seem like we are ahead of the problem of AI regulation. It is one thing to draft regulatory provisions for frontier technologies and another thing to have them be successful. One could argue that it is too early to judge the effectiveness of the aforementioned legal documents and their technical counterparts. Admittedly, the AIA entered into force less than a year ago, on August 1st, 2024, while most of the US legislative initiatives are still at a nascent stage, discontinued or not even voted into existence yet. Or has their time of evaluation already passed?

5.1 American - Chinese Arms race

Numerous observers have pointed out that there is an ongoing "AI arms race" between the United States and China, framing it as a pivotal competition that could shape the 21st century (Meacher, 2023). Both nations are pursuing strategies to dominate AI development. The sense of urgency is increasing, as AI is seen as a transformative technology with the potential to redefine economic, military, and scientific power.

Most would say that the U.S., with the early development of ChatGPT, had taken an early lead in developing large language models (LLMs), driven by venture capital and private sector innovation. Traditionally, the U.S. excels in frontier technologies, driven by a culture of innovation, access to global talent, and a competitive private sector. Major tech companies like Microsoft, Alphabet, Amazon, and Apple have absorbed or heavily invested in smaller AI startups, creating a robust ecosystem. Amazon has repeatedly invested billions into Anthropic, the startup company behind Claude AI[24], with competitors following suit. Meta has bought a couple of AI and data startups, aiming at even more acquisitions[25] [26] [27]. But even the startups that are seemingly operating autonomously rely heavily on the infrastructure of Big Tech.

Aiding its free market, the U.S. government has imposed export restrictions on advanced semiconductors, in an attempt to limit China's access to cutting-edge computing power[28]. However, it is evident that U.S. regulatory efforts lag behind the rapid pace of AI innovation, leaving much of the development to the private sector. As previously discussed, AI regulation in the United States of America is fragmented, inconclusive and lately oriented towards enhancing AI development to establish its international domination on this technological frontier rather than regulating it and guiding it responsibly.

China's AI strategy on the other hand is state driven, with a focus on national security, economic competitiveness, and societal management. The government heavily subsidizes AI research, semiconductor development, and computing power, while also imposing strict ideological controls on AI systems.

China leads in AI journal publications, patents, and the production of top AI researchers (Hart,2024)[29]. China has an edge in applying AI to industrial processes, scientific research, and manufacturing, leveraging its vast talent pool and government support. Western insiders however, considered that its restrictive policies on data access and ideological compliance hinder generative AI progress. Despite these challenges, Chinese companies like Baidu and Alibaba are gaining commercial traction with their AI models.

---

[24] Amazon Invests $4 Billion In Anthropic: A Paradigm Shift In AI
[25] Meta In Talks to Buy Korean AI Chip Startup Founded By Samsung Engineer
[26] Microsoft pays Inflection $650 mill in licensing deal while poaching top talents, source says | Reuters
[27] Microsoft partners with Mistral in second AI deal beyond OpenAI | The Verge
[28] US imposes export controls on chips for AI to counter China
[29] China Leads Global Generative AI Patent Race, UN Report Says — The U.S. Comes A Distant, Distant Second

Lately, DeepSeek, a Chinese startup, caused waves internationally when it launched its DeepSeek-R1 model in January 2025[30]. Its performance was on par with the latest model from OpenAI and it shocked the tech market when it was revealed that the cost of development was just a fraction of that of its American counterparts[31]. Developments like DeepSeek's success have put to sleep false beliefs that China's state-driven model, while effective in mobilizing resources and achieving hard metrics like patents and publications, faces significant challenges due to its restrictive policies on data and ideological compliance. These constraints were supposed to limit China's ability to compete in generative AI, where openness and creativity are paramount. China's focus on industrial and scientific applications of AI has positioned the superpower as a potential leader in fields like robotics, healthcare, advanced manufacturing and now generative AI.

The reaction to DeepSeek's success from the markets and its competitors was immediate. As it required far less computational power to be trained, it resulted in a 17% drop for NVIDIA's stock price, wiping out 589 billion USD of market value, as it challenged the notion that AI requires extraordinary computational power to be developed[32]. Perhaps even more interesting and supportive of the narrative suggesting that an arms race is under way, are the remarks voiced by OpenAI towards the US Office of Science and Technology Policy[33].

The document issued by OpenAI is highly ideologically charged, something unexpected for a document published by a tech company. It mentions that PRC and namely the Chinese Communist Party, wishes to build «autocratic, authoritarian AI». OpenAI supports that «there is significant risk in building on top of DeepSeek models in critical infrastructure and other high-risk use cases given the potential that DeepSeek could be compelled by the CCP to manipulate its models to cause harm. And because DeepSeek is simultaneously state-subsidized, state-controlled, and freely available, the cost to its users is their privacy and security, as DeepSeek faces requirements under Chinese law to comply with demands for user data and uses it to train more capable systems for the CCP's use. Their models also more willingly generate how-to's for illicit and harmful activities such as identity fraud and intellectual property theft, a reflection of how the CCP views violations of American IP rights as a feature, not a flaw. ». To combat this

[30] DeepSeek-R1 Release | DeepSeek API Docs
[31] DeepSeek: The Chinese AI Startup Reshaping the U.S. Tech Industry
[32] Nvidia Loses $589 Billion as DeepSeek Batters Stock: Evening Briefing Americas - Bloomberg
[33] OpenAI's proposals for the U.S. AI Action Plan | OpenAI

national security threat OpenAI calls for «voluntary partnership between the federal government and the private sector», deregulation, as AI regulations will slow down innovation, export controls to «restrict the flow of AI technologies to PRC»

The AI arms race between the U.S. and China represents a critical juncture in global technological and geopolitical competition. As previously discussed, the latest regulatory signs from the US aren't the most assuring. The present administration seems more focused on making sure that China doesn't outperform the American companies rather than making sure that the systems developed are safe. As stated by Yoshua Bengio, computer scientist, and a pioneer of artificial neural networks and deep learning, «the effort is going into who's going to win the race, rather than how do we make sure we are not going to build something that blows up in our face»[34]. The common outcome of every arms race is corner-cutting when it comes to ethics and safety. And although effective regulatory frameworks to prevent catastrophic outcomes would be advantageous for both the USA and PRC, the singular focus on «getting there» and not on how to responsibly get there, leads to a complete dismissal of the efforts to create such frameworks.

5.2 **Where is Europe at?**

Europe's leadership in consumer protection, particularly through the General Data Protection Regulation (GDPR), has set a global benchmark for data privacy. Lately, the AIA has tried to emulate these previous achievements on the AI regulatory front (Pham, Davies, 2024, p.12). However, the region's well-intentioned regulations are increasingly seen as a double-edged sword, creating uncertainty and stifling innovation in the AI sector. This regulatory environment has led to delays in AI product launches, with companies wary of potential penalties and fines. The resulting hesitation from businesses has created a vicious cycle, where cautious regulators and hesitant companies leave European consumers and businesses at a disadvantage. Some experts have even suggested that companies may also be making a political statement about overregulation, further complicating the landscape.

Despite regulatory barriers, the EU's single market is simply too large to ignore with European consumers demonstrating strong demand for AI tools. Experts agree that clarity and

[34] https://news.sky.com/story/ai-arms-race-risks-amplifying-the-existential-dangers-of-superintelligence-13303309

collaboration are essential. Regulators must provide clearer guidelines to reduce legal uncertainty, while companies must engage proactively to shape policies that balance innovation with protection. The EU AI Act, if implemented wisely, could set a global standard for responsible AI. However, if it becomes overly restrictive, it risks deepening Europe's technological lag.

Indicative of the EU's position, stuck in between innovation and regulation are the remarks made by Sam Altman, Open AI's CEO and widely considered among the most important people in AI development. Altman, as part of a panel discussion held on the Technische Universität Berlin[35], expressed his belief that most Europeans wish to see AI developed and deployed within Europe to revitalize economic growth and advance scientific progress. He conveyed OpenAI's aspiration to establish a European counterpart to the *Stargate* project ("Stargate Europe"), emphasizing the necessity of support for such an initiative, which would be governed and operated within the European framework.

He underscored the importance of ensuring that OpenAI's products become readily available in Europe, just as they are in other parts of the world. He affirmed his view that a strong Europe is crucial for global stability and expressed confidence that Europeans would excel in leveraging AI technology. Commenting on the AIA, Altman stated that OpenAI would comply with European legislation and respect the will of the European public. He acknowledged that Europeans would determine the regulatory framework, and OpenAI would operate within those established rules.

Furthermore, he advocated for European access to cutting-edge AI technologies, arguing that a lack of access would confer no strategic advantage. He maintained that the only way to shape technological development is through direct engagement—by working with the technology, understanding it, and innovating within it. While not opposing regulation in principle, he cautioned against excessive restrictions, particularly at this early stage of AI's evolution. Altman also raised concerns about whether Europe's top AI talent would focus primarily on regulatory compliance or on seizing opportunities and solving complex problems.

Overall, Altman remains optimistic about AI's potential in Europe and supports close collaboration. While he affirms his commitment to regulatory compliance, he subtly signals a preference for a regulatory environment that fosters innovation rather than stifling it, ensuring that Europe fully capitalizes on AI's transformative potential.

[35] https://www.youtube.com/watch?v=McuO7Osgzqo

Given Europe's position, the AI summit held in Paris, France, on February 2025, acts as a precursor of what's to come for the continent's AI development efforts[36]. Among the speakers at the summit, the one with the most authority to speaks of EU's future was the President of the European Commission Ursula von der Leyen.

President von der Leyen emphasized Europe's aspiration to establish itself as a leading continent in the field of Artificial Intelligence (AI), integrating AI into every facet of daily life. She underscored the manifold benefits that AI can offer to Europe, including enhancing competitiveness, safeguarding security, strengthening public health, and facilitating access to knowledge.

Refuting the claim that Europe has lagged behind in the global AI race, she asserted that the race has only just begun, and that global leadership remains attainable. The European approach, she highlighted, focuses on embedding AI within specific industrial applications and leveraging it for productivity and human-centric advancements. President von der Leyen delineated the key characteristics of European AI. She claimed that the focus is on complex applications, prioritizing the adoption of AI in sophisticated use cases, capitalizing on Europe's unique industrial and manufacturing data, as well as its technical expertise, via assembling talent from various nations and sectors, building upon the tradition of European scientific cooperation and advocating for open-source AI as a mechanism for accelerating technological dissemination.

To further the development of European AI, she announced targeted measures aimed at startups and researchers, granting them access to European supercomputers for the development and testing of their models. Additionally, she unveiled the establishment of 12 "AI factories" with a public investment of €10 billion—described as the largest public investment in AI globally—with the aim of attracting significantly greater private investment and ensuring computational accessibility for all enterprises. Concurrently, she introduced the concept of "AI Giga-factories," large-scale data and computing infrastructures inspired by the success of CERN.

Emphasizing the necessity of cooperation and trust in AI, as well as its safety, she announced the introduction of the "AI Act," which will establish a unified framework of security regulations applicable across all EU Member States while simultaneously reducing bureaucratic obstacles.

[36] https://www.youtube.com/watch?v=wxLnHa-7qls

Addressing the financial demands of AI innovation, she welcomed the "European AI Champions" initiative, which entails commitments amounting to €50 billion, and introduced the "Invest AI" initiative, allocating an additional €50 billion. These efforts aim to mobilize a total investment of €200 billion in European AI, with a particular emphasis on industrial applications, thereby constituting the largest public-private partnership for the development of trustworthy AI. Furthermore, she reaffirmed support for the "AI Foundation" and articulated a vision for AI as a force for the common good, ensuring that its benefits remain universally accessible.

Of equal interest was the address of Finland's Prime Minister, a European country traditionally considered to be at the forefront of technological innovation in the continent. Following the summit, France's President Emmanuel Macron declared that a 109 billion euros investment will be made to help AI development on an EU level[37]. While this might sound like a lot of money, it is just a fragment of what the US's private and public sector are willing to raise for AI research and development. Apple has recently committed 500 billion US dollars to AI research, while President Donald Trump has announced the Stargate Initiative, a $500 billion private sector deal to expand U.S. artificial intelligence infrastructure[38].

President von der Leyen seems to signal for a new age of Artificial Intelligence research and development in Europe, sounding hopeful that the Union isn't forever behind the already substantially advanced USA and PRC. The recent success of DeepSeek proves that AI success is feasible without the mammoth investments that take place in the US. However, as previously discussed, one of the main driving forces in the US is the high-tech ecosystem created by the big tech companies which, being in an internal competition with each other, in addition to the international competition they now face, will most likely be more than unwilling to defer their infrastructure into European AI projects, given the added restrictions of AIA.

Finally, European endeavours will have to be carried out strictly within the regulatory framework set by AI Act. The concerns about the Regulation's applicability should be taken seriously. The Regulation might nobly require extensive technical documentation, transparency of AI systems and automatic record keeping, but it is yet to be proven that such requirements can be technically met. Otherwise, they might simply end up putting undue burden on the European AI programs.

[37] Macron signals investments of 109 billion euros in French AI by private sector | Reuters
[38] Stargate: Trump announces a $500 billion AI infrastructure investment in the US | CNN Business

5.3 Those who innovate don't regulate and those who regulate don't innovate

It is evident that the EU has put forth the most comprehensive AI regulation to date by putting into force the AI Act, but is currently miles behind the main AI players in the USA and PRC. While the latter have no to little AI regulatory provisions in effect. USA's acting administration is more focused to using the executive branch of the government to enable AI RnD without much concern about trying to uphold any of the values that regulators hold so dear, echoing Silicon Valley's famous attitude of move fast and break things, solely focused on dominating its international competition, namely China.

This arms race raises important questions about the future of AI governance and the need for international cooperation. The development of global AI safety frameworks is essential to mitigate risks such as the misuse of AI technologies. Both the U.S. and China have a vested interest in ensuring that AI advancements benefit humanity as a whole, rather than exacerbating geopolitical tensions. Collaboration on global AI safety and governance could provide a pathway for mutual benefit, ensuring that AI remains a force for progress rather than a source of conflict. The choices made today will shape the trajectory of AI development and its impact on society for decades to come.

5.4 International Cooperation

AI governance frameworks must be designed to ensure interoperability across different jurisdictions, rooted in universally recognized principles such as those outlined in the Universal Declaration of Human Rights. To achieve this, existing international institutions like the ITU should be utilized to strengthen the alignment of regulatory measures globally (UN, 2023, p.16). Furthermore, the establishment of a centralized body could facilitate the harmonization of policies, foster shared understandings, highlight best practices, assist in implementation, and encourage peer-to-peer learning among stakeholders (UN, 2023, p.13).

A Global AI Governance Framework should be developed to provide a cohesive structure for policymaking and implementation, addressing potential disparities in AI governance between public and private sectors, regions, and nations. This framework would also clarify the foundational principles and norms guiding the operations of various organizations. Special emphasis should be placed on capacity-building initiatives for both private and public sectors,

alongside efforts to disseminate knowledge and raise awareness on a global scale. Best practices, such as conducting human rights impact assessments for AI systems, should be promoted through this framework, potentially requiring an international agreement to ensure widespread adoption and compliance (UN, 2023, p.14).

Such proposals underscore the necessity of a harmonized approach to AI governance, which is critical in an era where technological advancements often outpace regulatory frameworks. By anchoring governance arrangements in established international norms like the Universal Declaration of Human Rights, the framework ensures that ethical and legal considerations remain at the forefront of AI development and deployment. Leveraging existing UN bodies such as UNESCO and the ITU is a pragmatic approach, as these organizations already possess the institutional infrastructure and expertise to facilitate cross-jurisdictional cooperation.

The creation of a global body to coordinate AI governance efforts is a forward-thinking recommendation. Such a body could serve as a central hub for policy harmonization, knowledge-sharing, and capacity-building, addressing the current fragmentation in AI governance. The emphasis on capacity-building is particularly noteworthy, as it acknowledges the disparities in technological and regulatory expertise between developed and developing nations. Ensuring that all countries have the tools and knowledge to participate in AI governance is essential for achieving equitable outcomes.

However, the success of such measures will depend on the willingness of states and private entities to adopt and enforce these practices. The call for an international agreement to underpin the Global AI Governance Framework is both ambitious and necessary. While achieving consensus on such an agreement may be challenging, it is a critical step toward ensuring that AI governance is not only interoperable but also universally respected and implemented.

G7 members have too emphasized the necessity of establishing robust regulatory frameworks and ensuring interoperability across jurisdictions to govern AI effectively (OECD, 2023, p.23). One member highlighted the importance of building on existing frameworks, such as the OECD AI Principles, which call for international norms, standards, and assessment processes. This member advocated for global alignment and collaboration, particularly including developing countries to ensure inclusivity.

Another member underscored the critical role of international governance, including the creation of common governance frameworks and international standards for the reliability of generative AI (e.g., quality control). They emphasized the need for international institutions to enable rapid, coordinated responses to emerging and unknown threats. This member drew parallels to other areas of global governance, such as climate change and international trade, where codified and institutionalized cooperation has proven beneficial (OECD, 2023, p.23).

A different G7 member stressed the importance of developing precise and detailed principles to facilitate their implementation within G7 countries. They suggested that these principles could be operationalized through general agreements with generative AI system providers or, if necessary, through binding legislation to ensure compliance (OECD, 2023, p.23).

Other members highlighted the need for greater interoperability of regulatory frameworks across jurisdictions, as well as the development of common guidelines to promote responsible AI development and use.

The responses from G7 members reflect a growing recognition of the complexities and global nature of AI governance. The call for international alignment and collaboration, particularly with developing countries, is a crucial step toward ensuring that AI governance is inclusive and equitable. Building on existing frameworks like the OECD AI Principles provides a solid foundation, as these principles are already widely recognized and incorporate key ethical and technical considerations.

The emphasis on international governance and the establishment of common frameworks for generative AI reliability is particularly timely, given the rapid advancements in this field. The comparison to other global governance areas, such as climate change and trade, is apt, as these domains demonstrate, however, certain deficiencies of institutionalized cooperation in addressing transnational challenges. Couple that with the dynamic and evolving nature of AI technology presenting unique challenges, necessitating flexible yet robust governance mechanisms and we have a clear outline of the challenges we need to face in order to successfully implement such a global framework.

The suggestion to implement principles through general agreements with AI providers or binding legislation is a pragmatic approach. While voluntary agreements may foster collaboration and innovation, binding legislation may be necessary to ensure accountability and

compliance, particularly in high-stakes applications of AI. This dual-pronged approach allows for adaptability while maintaining a baseline of enforceable standards.

The focus on interoperability of regulatory frameworks is essential to avoid fragmentation and ensure that AI systems can operate seamlessly across borders. Common guidelines for responsible AI development and use are equally important, as they provide a shared understanding of ethical and technical standards, reducing the risk of misuse or harm.

The G7 members' perspectives and UN positions highlight the need for a coordinated, inclusive, and adaptable approach to AI governance. By leveraging existing frameworks, fostering international collaboration, and ensuring interoperability, the global community can address the challenges posed by AI while maximizing its benefits. However, achieving this will require sustained political will, technical expertise, and a commitment to balancing innovation with ethical and legal safeguards, all of which are not guaranteed when it comes to Artificial Intelligence.

6. CONCLUDING THOUGHTS. CAN WE REGULATE OUR WAY TO RESPONSIBLE AI?

There is one inescapable truth. Artificial intelligence is a part of our lives and it is here to stay. Everyone is in awe of its possibilities. Many consider it to be the solution to all our problems. The invention to end all inventions. Others worry it will spell humanity's inevitable doom. Sam Altman himself has said that «I think, AI will probably, most likely sort of lead to the end of the world» and he is not the only one to characterize AI as an existential threat to humanity. The variety of outcomes is, to say the least, extreme.

In the face of such great stakes, many have tried to determine how to avoid the abyss and lead humanity to Eden instead. Among experts internationally, a consensus seems to have formed that in order to prevent a catastrophe, we not only need to create AI systems but make sure that we design them as responsible AI systems. To achieve this, the AI systems we create need to be transparent, accountable, fair and unbiased, respect privacy, robust, secure and under some degree of human control.

A couple of steps have been made towards incorporating those principles in AI research and development. Most notably, the European Union put into force Regulation (EU) 2024/1689 of the

European Parliament and of the Council of 13 June 2024, commonly referred to as «The Artificial Intelligence Act (AIA). This Regulation as outlined above, performs a trichotomy of AI systems based on the risk their operation would pose, banning the ones that pose unacceptable risk, imposing detailed and extended requirements for the systems that are deemed high-risk and less strict measures to the systems that present low or minimal risks.

However, as we saw, the EU is nowhere to be found yet when it comes to AI systems development globally. Just in the beginning of 2025, EU officials and leaders of prominent member states announced funding for serious AI research hoping to catch up to the other superpowers who have already made leaps in the field. Therefore, while the EU might have presented a comprehensive Regulation aiming at safeguarding the creation of responsible AI systems, the footprint of the Regulation is minimal as most of the AI developments take place outside the Union. Of course, the Regulation applies to AI systems that, while developed outside the Union, are put into service within it. This stipulation too hasn't seemed to affect at all the way US and Chinese companies develop their systems as they are just not that concerned with introducing them to the EU market, preferring to restrict certain features to non-EU based users.

On the other hand, the AI innovators are for the time being the United States of America and the People's Republic of China. US and China based companies such as OpenAI, Anthropic, DeepSeek, Alphabet Inc. and xAI lead the way, seemingly introducing new and updated AI systems every week. Not only are these companies heavily incentivised to not only outperform their competitors but to do so faster than the rest of the field, securing immense economic benefits, but so are the countries. Both the White House and Beijing have announced AI research and development programs, providing funding and further incentives, hoping to be the ones to achieve the next significant breakthrough and thus cementing themselves as the leading force in AI systems, securing substantial strategic advantages.

This predisposition reflects in the regulatory decisions made. The current US administration has sidelined most of the previously proposed bills and executive orders, aiming to emulate the EU's path towards responsible AI systems, and has instead favoured administrative and legislative actions aiming to enable US-based companies to move faster and with fewer obstacles in their way to the next big AI – related breakthrough, while trying to stifle their Chinese counterparts. The little meaningful legislation that exists in the US is at mere state level, not able to influence federal policy making. Furthermore, the technical guidelines, aligned with the generally accepted

principles for Responsible AI, are not binding and enforceable upon corporations to be able to play a decisive role, unless they are voluntarily implemented.

It is becoming all the more evident that an arms race is developing between the USA and China, regarding AI systems. Such a dynamic would mean that companies are paying less attention to the ethics and the safety surrounding their AI programs, focusing solely on their effectiveness. Couple that with the deregulation trend adopted by the US administration and the conditions are ripe for a misstep with far-reaching effects. Such ramifications will most likely exceed national borders.

Subsequently, the problem at hand emerges as a global one and only a global solution can be effective. The efforts to establish Responsible AI principles need to be cooperative and coordinated at an international level, much like the measures to prevent an ecological catastrophe need to be horizontally adopted and enforced and not by just a couple of eco-friendly countries. A coordinated international governance framework rises as the one effective option to combat the status quo in AI development. By enforcing the same set of rules to every AI system developer, we can minimize the risk of unwanted outcomes materializing, as the arms race dynamic would be alleviated, or at least take place in a regulated environment.

However, as previously stated, such a resolution will require, first and foremost, sustained political will. Not by just one administration, as the EU has shown, but by all the main geopolitical players. And this political will, must converge, from opposing starting points of self-interest, to a commitment of balancing innovation with ethical and legal safeguards. Such political will doesn't seem to be available at the moment of the writing of this thesis.

Moreover, even if the agreement was reached to horizontally and internationally regulate the development of artificial intelligence systems, the technical aspect is still not clear. The AIA could be used as a building block for an international regulation, but its applicability is still to be proved. The pillars of responsible AI, such as transparency and human oversight, might not be technically attainable given the dynamic and intricate nature of AI systems. And even if they are achievable at the current state of AI systems, we cannot be certain that they will be moving forward.

The subject of this thesis has been Responsible AI. Its paramount importance is undermined by the grave dangers posed by the development and application of AI systems. To avoid these risks,

a set of principles has emerged as the answer. Transparency, safety, robustness and human oversight are just a few among them. Attempts to enforce this set of principles have been made, most notably by the EU and the adoption of the AI Act. However, the current state of AI development, a field dominated by the US and the PRC, two powers seemingly competing to reach AI dominance. This climate greatly undermines the safety discourse and any international agreements that could be made regarding Responsible AI. Coupling the international political climate with the yet unanswered questions regarding the actual technical attainability of Responsible AI leaves us with little to comfort ourselves with about what the future holds. Responsible AI might just turn out to be the ultimate regulatory wishful thinking.

REFERENCES

Agrawal, A., Gans, J., & Goldfarb, A. (2022) *"Prediction Machines: The Simple Economics of Artificial Intelligence"*, Harvard business Review Press

Akkermans, H. (2024) *'The Social Responsibilities of Scientists and Technologists in the Digital Age'*, in Nida-Rümelin, J., Nuseibeh, B., Prem, E., Stanger, A. (eds) Introduction to Digital Humanism. Springer.

Bostrom, N. (2014), *"Superintelligence, Paths, Dangers, Strategies"*, Oxford University Press

Brey, P. and Dainow, B. (2021) *'Ethics by design and ethics of use in AI and robotics', The SIENNA project-Stakeholder-informed ethics for new technologies with high socioeconomic and human rights impact.* European Commission

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., Lee, P., Lee, Y.T., Li, Y., Lundberg, S., Nori, H., Palangi, H., Ribeiro, M.T., & Zhang, Y. (2023) *'Sparks of Artificial General Intelligence: Early experiments with GPT-4'*. Microsoft Research.

CAIDP (2023) OPEN AI (FTC 2023) Complaint. Center for AI and Digital Policy. Available at: https://www.caidp.org/cases/openai/

Card, S. K., (1983) *"The psychology of human-computer interaction".* CRC Press.

Coglianese, C. & Ben Dor, L. M. (2021*) 'AI in Adjudication and Administration'*, Brooklyn Law Review, 86, p. 791.

Crawford, K. (2021) *Atlas of AI: Power, Politics, and the Planetary Costs of Artificial Intelligence.* Yale University Press

Crootof, R., Kaminski, M. E. & Price II, W. N. (2023) *'Humans in the Loop'*, Vanderbilt Law Review, 76, p. 429.

Cummings, M. L., Roff, H., Cukier, K., Parakilas, J., Bryce, H. (2018) *Artificial intelligence and international affairs: disruption anticipated.* Chatham House. Available at:

https://www.chathamhouse.org/sites/default/files/publications/research/2018-06-14-artificial-intelligence-international-affairs-cummings-roff-cukier-parakilas-bryce.pdf

European Commission (2021) *"IMPACT ASSESSMENT Accompanying the 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"*, European Commission

Felzmann, H., Fosch-Villaronga, E., Lutz, C., Tamò-Larrieux, A. (2020) *"Towards Transparency by Design for Artificial Intelligence"*, Springer

Gillis, R., Laux, J. and Mittelstadt, B. (2021) *'Trust and Trustworthiness in AI'* in: Paul, R., Carmel, E. and J. Cobbe (eds.), Handbook on Artificial Intelligence and Public Policy, Cheltenham Spa: Edward Elgar.

Harari, Y. N., (2017) *"Homo Deus: A Brief History of Tomorrow"* Harper

High-level Expert Group on Artificial Intelligence set up by the European Commission (2019), *"ETHICS GUIDELINES FOR TRUSTWORTHY AI"*, European Commission

International Organization for Standardization (2023) ISO/IEC TR 23894:2023, Information technology — Artificial intelligence (AI) — Guidance on risk management. ISO/IEC.

International Organization for Standardization (2023) ISO/IEC 42001:2023, Information technology — Artificial intelligence — Management system. ISO/IEC.

Kaadoud I. C., Fahed L., Lenca P., (2021), *"Explainable AI: a narrative review at the crossroad of Knowledge Discovery, Knowledge Representation and Representation Learning. MRC 2021: Twelfth International Workshop Modelling and Reasoning in Context"*, Montréal (virtual), Canada. pp.28-40. hal-03343687

Kökciyan, R., Broeders, J., Wintraecken, A., Tazelaar, A., Summerfield, C., Binz, M., Hellwig, O., Zylberberg, T., Carey-Galea, L. and Turchin, P. (2022) *"Human-centred mechanism design for democratic AI"*, Nature Human Behaviour.

Leslie, D., Rincón, C., Briggs, M., Perini, A., Jayadeva, S., Borda, A., Bennett, S.J., Burr, C., Aitken, M., Katell, M., Fischer, C., Wong, J. and Garcia, I.K. (2023) *"Fairness by Design: A Guidebook for Integrating Fairness into the Development of Data-Driven Systems"*. The Alan Turing Institute.

Matsumi, H., Solove D. J., (2024), *"The Prediction Society: Power, Prediction, and the Law"*.

Microsoft (2023) *"Governing AI: A Blueprint for the Future"* Microsoft

Mikalef, P., Conboy, K., Lundström, J. E., Popovič, A., (2022) *"Thinking responsibly about responsible AI and 'the dark side' of AI"*, European Journal of Information Systems

Nida-Rümelin, j., Winter D. (2024) *"Humanism and Enlightenment"* in Nida-Rümelin, J., Nuseibeh, B., Prem, E., Stanger, A. (eds) Introduction to Digital Humanism. Springer, Cham.

OECD «HIROSHIMA PROCESS International Guiding Principles for Organizations Developing and Using AI», 2023

OpenAI (2025) *"OpenAI's proposals for the U.S. AI Action Plan"*, OpenAI

Pham, B.C. and Davies, S.R. (2024) 'What problems is the AI act solving? Technological solutionism, fundamental rights, and trustworthiness in European AI policy'.

Pitoura, E. (2020) 'Social-minded Measures of Data Quality: Fairness, Diversity, and Lack of Bias', ACM Journal of Data and Information Quality, July.

Van der Sloot, B., (2024) Regulating Synthetic Society: The Law and Ethics of AI, Deepfakes, Augmented Reality and Virtual Reality. Hart Publishing.

Wörsdörfer, M., (2023) «Mitigating the adverse effects of AI with the European Union's artificial intelligence act: Hype or hope?», Wiley Periodicals LLC.