POLITECNICO DI TORINO

Master's Degree course in Cybersecurity

Master's Degree Thesis

Technique, Trust, and Transparency: A Critical Examination of the EU AI Act in the Age of the Technological Society

Advisor
Prof. Giuseppe Emiliano Vaciago

Co-Advisor
Prof. Juan Carlos De Martin

Candidate
Anna Alexandra Antonini

Academic Year 2024-2025

Acknowledgements

My sincere appreciation goes to the Professionals of the LT42 Law Firm. I am deeply indebted to them for the exceptional opportunity to work alongside them, gathering information and practical experience for this thesis. Their graciousness in welcoming me into their practice provided a vital, real-world context for my research.

I am particularly grateful for the intellectual generosity of Professor Giuseppe Vaciago's Partners and Associates, especially Mr. Gianluca Gilardi, Ms. Martina Elisa Mauloni, and Mr. Michele Pellerzi, who patiently addressed my queries and shared their views on this emerging regulatory landscape. The insights gained from my experience were instrumental in shaping the core arguments of this dissertation and have been invaluable for both my professional and personal growth.

Chapter 1

Introduction

This Master's Thesis examines the rise of Artificial Intelligence as both a technology and a shaping force that, despite its relative youth, has already had a significant impact on human life and society, creating an urgent need for regulation worldwide. This work wants to contribute to the global discussion by integrating relevant academic and news sources with realistic case analysis, informed by my experience at LT42 headquarters in Milan. The analysis is grounded in the theoretical framework of Jacques Ellul and other contemporary thinkers, placing the AI revolution within a broader discourse on society and its relationship with technology.

This thesis is structured around three core, interconnected **research questions**, which guide the comparative analysis of the EU's approach to AI governance:

- How does the EU AI Act balance the imperatives of innovation and fundamental-rights protection within its structure, and how effectively do its mechanisms support this balance?
- What are the benefits and limits of the Act's risk-based classification when applied to opaque, general-purpose AI models (GPAI), and how do the black-box nature of these systems challenge the need for transparency, robustness, and accountability?
- How does Jacques Ellul's *The Technological Society* help explain the public attitude towards AI?

Methodology note A multifaceted methodological approach combines legal textual analysis, comparative case studies, and literature review to evaluate the European Union's AI Act (AIA) and its theoretical underpinnings.

The research that has gone behind it is primarily focused on dissecting the current status of affairs to better understand the context and the practical challenges behind a correct implementation of the Act.

To provide a real-world context and illustrate the practical application of the AIA's framework, specific case studies were developed to demonstrate the real-life tensions that can arise in the practice of a computer engineer who needs to interface with the law governing their work. The legal analysis is cross-referenced with technical research and external data to ground normative concepts in empirical reality, such as technical

standards, quantitative metrics proposed by academic work, and reports on enforcement actions by EU Data Protection Authorities against major AI providers.

This methodology aims to provide a comprehensive picture of the status quo, addressing both the letter of the law and its operation within the complex, rapidly evolving landscape of the modern Technological Society.

Part I

Comparative Overview

Chapter 2

International and Multilateral AI Governance Initiatives

Governments are increasingly striving to keep pace with the rapid evolution of technology and its accelerated expansion into markets and the public sphere. A new wave of laws and policies is emerging worldwide to address the impact of technology on citizens' lives and work:in 2024 alone, references to AI in legislation increased by over 21% globally [110], representing a ninefold rise since 2016.

Reflecting the general-purpose nature of this technology, the challenges associated with AI are diverse, ranging from established concerns such as **data privacy** to emerging issues regarding the attribution of **responsibility** for the actions of autonomous algorithms.

It is important, then, to provide at least an overview of key AI laws across major countries and regions, to highlight regional differences and international coordination efforts.

- In 2022, UNESCO's *Recommendation on the Ethics of AI* [122] represented a landmark in international AI policy as the **first global ethical standard for AI** [121]. This framework centres on the **protection of human rights and dignity**, establishing key principles including **transparency**, **fairness**, **accountability**, **privacy**, and **human oversight** of AI systems. It further provides policy **guidelines** to operationalise these principles, particularly in areas such as data governance, education, culture, and environmental issues.

- In November 2023, the United Kingdom hosted the *Global AI Safety Summit* at Bletchley Park[117], an event that brought together representatives from countries including Australia, China, Israel, Nigeria, the Republic of Korea, Rwanda, Switzerland, the United Kingdom, and the United States, as well as the European Union. The summit addressed risks associated with highly advanced AI systems that could cause **serious or catastrophic harm**, whether intentional or accidental, particularly in critical sectors such as **cybersecurity** and **biotechnology**. The discussions resulted in the *Bletchley Declaration*, which emphasised the urgency of addressing these risks. Notably, Russia, despite its relevant role in the industry,did not participate in the summit.

- In a further multilateral initiative, the **Council of Europe**, comprising 46 countries, is currently drafting a **Framework Convention on AI, Human Rights, Democracy, and the Rule of Law** [63], with contributions from notable participants such as the United Kingdom, the United States, and Israel. The convention seeks to establish common standards for AI governance, accountability, and risk assessment consistent with human rights norms.

These international initiatives indicate a **growing consensus** on high-level AI principles, such as transparency, accountability, and human oversight. However, in practice the degree and manner of **implementation** differ markedly from one jurisdiction to another, reflecting variations in local **priorities**, regulatory approaches, and levels of technological development. For example, while some countries have adopted comprehensive regulatory frameworks with binding legal requirements, others rely primarily on non-binding guidelines or sector-specific measures. These differences underscore the complexity of achieving international alignment on AI governance, between shared principles and divergent strategies.

2.1 International AI standards and agreements

This section provides a brief review of what brings actors together,namely, the active **standards** in the field of AI, and their role at this stage in the evolution and development of artificially intelligent technologies.

2.1.1 The importance and role of standard setting in technology

In the modern geopolitical landscape, standard-setting has become a key arena for **competition**, particularly among major powers such as the United States, Europe, and China, each adopting distinct strategies to **shape** the rules governing emerging technologies like AI. Standards in technology, especially if it is a new, budding field of it, are not merely about bureaucratic control: they are a **strategic necessity** for enabling domestic **innovation**, ensuring **market predictability**, reliability and **interoperability**, at the same pace of building **trust** in its safety. They serve as vital tools for effective governance and risk management, and promote transparency and ethical development. Their role as valid guidelines to follow for organisations is widely recognised, with reports [66] of AI standards allowing for fewer issues after deployment and **faster**[18] time to market for new products. Indeed, standards have always acted as a common language that enhances mutual understanding on the global panorama (with estimates[95] suggesting an impact on about 93% of global trade): they provide clear **benchmarks** for new products entering the market, promote fair **competition** against monopolies[1], and allow for new **interconnections** for integrated systems that are more than the sum of their parts.

As can be suspected, with such significant involvement comes also the **power** to be gained for those who dictate the rules. In countries were state control tends to bow

[1]See, for example, the case of the USB-C standard, that allowed for proliferation of third-party peripherals and reuse of old charging cables from different manufacturers.

to industry power, such as the United States, the process of standard-setting is born out of **self-organized** conferences between market leaders, which then pass on the proceedings to government organisations for final approval. This is a particularly profitable arrangement for the companies that end up on top, as they are set to gain from intellectual property **licensing** and, indirectly, from locking incompliant firms out of Western markets. On the other side of the spectrum, countries like China have a strong top-down approach that signifies strong **government involvement** , which reinforces the common belief that standard setting is a prerogative of world-leading players[17]. Indeed, one of the Chines Government's primary objectives is to free Chinese industry from paying expensive royalty fees for foreign-held patents embedded in international standards, while simultaneously establishing itself as the market leader. Through backing coming from commercial agreements such as the *Belt and Road Initiative*, China has pushed[46] for international **alignment** with standards bodies it has particular influence over, like the United Nations' *International Telecommunication Union* (ITU), especially in the field of mobile communications[76].

As **China**'s approach[17, 25, 18, 46, 108] is to treat IP as a factor of production whose cost should be **minimised**, this has led to a push for cheap or **royalty-free** licensing options for patents included in standards, and to the adoption of standards as an effective leverage to negotiate better economic conditions (particularly in lowering royalty rates for Chinese firms). This strategy, combined with the creation of plausible alternatives, has allowed China to introduce competition multiple times into what would otherwise be a monopoly on essential technology, directly benefiting its vast manufacturing sector, even when the unique Chinese standards themselves failed to gain widespread adoption. A significant precedent for this is what happened with **Blu-Ray Players** in the early 2000s: at that time, Chinese factories produced about 75%[61] of the world's DVD players, but their profit margins fell to as low as one U.S. dollar[16] per unit due to high royalty payments to a consortium of foreign patent holders. In response to this, China's *Ministry of Information Industry* backed the development of an indigenous standard called EVD (Enhanced Versatile Disc). Foreign patent holders of the dominant standard were now faced with a choice: maintain their high royalty rates and risk Chinese manufacturers shifting to the new, cheaper domestic standard, potentially **losing access to the vast Chinese manufacturing ecosystem** and domestic market, or offer a significant reduction in royalty rates to make their own standard more attractive and maintain market dominance. This process was repeated in the high-definition era: China developed the China Brand High Definition (**CBHD**) disc as an alternative to **Blu-Ray**, which was a sufficient threat to force the Blu-Ray Alliance to not only include a Chinese audio-video codec in its international standard but also to reduce its mandatory royalties, despite the CBHD never reaching commercial success.

What is at stake is more than just profits, although companies such as Huawei, for example, now earn more from licensing its technology than it pays for others'[45], but a **conduit** for ideology, as standards can encode social values, just as it happened when Western-designed Internet was heralded as a bringer of edge user empowerment and decentralisation.

2.1.2 Main international AI standards and agreements

The AI **standards** landscape is populated by a variety of international and national bodies, each playing a distinct role in shaping the governance of this technology. Here is a brief list of the leading actors and their contributions:

ISO/IEC Their **ISO/IEC 42001:2023[69]** standard is considered the "gold standard" for an AI management system, providing a comprehensive framework for AI governance and risk management. Other publications include standards for *Artificial Intelligence concepts and terminology* (ISO/IEC 22989:2022) and *Guidance on risk management* (ISO/IEC 23894:2023)[68], and the tech reports on *Bias in AI systems and AI aided decision making* (ISO/IEC TR 24027:2021)[71], on ethical and societal concerns on the use of AI (ISO/IEC TR 24368)[73], on trustworthiness in AI (ISO/IEC TR 24028:2020)[72], and even on environmental sustainability concerns (ISO/IEC TR 20226:2025)[70].

IEEE The *Institute of Electrical and Electronics Engineers (IEEE)* has an active portfolio of standards related to autonomous systems, machine learning, and AI ethics, including the IEEE 2857-2024[64] standard for AI performance **benchmarking**.

ITU The I*nternational Telecommunication Union (ITU)* is actively developing guidelines for AI, which are formally published as *ITU **Recommendations***. In particular, the Union has dealt with the use of AI for 5G/6G communications (ITU-T Y.317x series[78]), for Health (ITU-T H.870, *Guidelines for safe listening devices/systems*[77]), multimedia and autonomous driving (the latter two being active, but still developing areas, involved in fields such as deepfake detection[75]).
Many of these contributions are part of the *AI for Good[74]* initiative, dedicated to ensuring Artificial Intelligence is used to help solve the world's most urgent problems, like climate change, hunger, poverty, health, and inequality, as outlined by the UN *Sustainable Development Goals* (SDGs), bringing together governments, private companies, academia, civil society, and 40 UN sister agencies on a single platform.

OECD,UNESCO The *Organisation for Economic Co-operation and Development* (OECD) and the *United Nations Educational, Scientific and Cultural Organization* (UNESCO) have established more high-level principles for ethical AI[122], as (non-binding) recommendations for Member States to implement them in their governing structures.

Chapter 3

The United States' approach: market-driven, sector-specific, and innovation-focused

The US regulatory landscape attempts to **balance** innovation and safety concerns in a relatively fragmented manner, characterised by **sector-specific** regulations addressing AI in particular domains, as well as various **state-level** initiatives that create variation across jurisdictions. On the one hand, executive orders provide federal-level guidance, but they lack the binding **force** of legislation. On the other hand, state-level action is generally relegated to **updating** and amending existing legal frameworks to account for the specific risks introduced by AI systems.

3.1 Data Privacy in the U.S.

Unlike the EU's GDPR, the U.S. lacks a blanket federal data protection law that covers the private sector's use of personal data [52]. However, sectoral privacy laws exist (for example, **HIPAA** for health data, **FERPA** for educational settings), together with a patchwork of state laws: California's Consumer Privacy Act (CCPA)[19], effective 2020, granted California *residents* rights over personal data and imposed obligations to *businesses* that meet certain thresholds. Later, in 2023, California's *Privacy Protection Agency* also presented draft regulations that to require businesses to disclose and offer **opt-outs** for **automated decision-making** systems that significantly affect consumers (such as AI used in job hiring or credit decisions)[20]. Several other states (Virginia, Colorado, Connecticut, Utah, etc.) have passed privacy laws, though these will not be analysed since they are generally weaker than California's [120].

3.2 Federal Initiatives and International Standards

At the *federal* level, a single, *general* AI legislation remains under debate, so the United States have yet to adopt a **centralised** approach to AI governance. However, collectively, the *AI in Government Act*, the *AI Leadership Order*, and the *Trustworthy AI Order* remain the three critical pillars of the complete U.S. strategy, in addition to relying heav-

ily on a combination of **existing** laws (such as consumer protection, anti-discrimination, and privacy laws), **agency guidance**, **state-level** initiatives, and **voluntary frameworks**[21].

In the meantime, federal bodies are using existing authority to oversee AI uses within their domain: the *Federal Trade Commission* (FTC), for example, treats biased or deceptive AI outputs as potential violations of consumer protection laws (as per the *FTC Act*, Section 5), the *Equal Employment Opportunity Commission* (EEOC) looks for compliance with civil rights laws on employment discrimination, and the *Consumer Financial Protection Bureau* (CFPB) has intervened on giving explanations to automated decisions under the *Equal Credit Opportunity Act*[51].

Generally speaking, the USA's stance on regulatory matters is always one that fosters innovation and growth, explicitly [89] [84] preferring «alternatives to regulation, and [...] public communications and [...] voluntary consensus standards, that agencies could take to reduce barriers to the use of AI.", rather than imposing a more precautionary approach [88]. Following this philosophy, in *lieu* of hard law, the U.S. government and its Offices and Bodies have issued many policy frameworks and executive actions on AI.The 2021 *National AI Initiative Act* established a **coordinated** federal strategy for *AI Research & Development* and set up the *National AI Initiative Office*, as well as the *National AI Advisory Committee* and the *National AI Research Resource Task Force*; as many other Offices and Committees before them on strategic and high-impact technologies, these efforts were focused on **funding research** and **technical education**, rather than regulation[131]. All this clearly reflects the USA's desire not to fall behind in innovation and retain **world leadership** through investment and standard-setting utilising their economic and political influence, rather than approaching carefully a yet scarcely understood technology and undermine its growth. In 2022, the *White House Office of Science and Technology Policy* published a *Blueprint for an AI Bill of Rights*[105], which contains a **non-binding** set of principles for the design and deployment of AI systems. This document calls for **safe** and **effective** systems, algorithmic **discrimination** protections, data **privacy**, clear **notice** and **explanation** about AI use, and human-based **fallback** plans. Though just advisory, it set the tone for subsequent governance efforts.

Figure 3.1: OECD's *Principles for trustworthy AI* [1]

As far as **standards** are concerned, in January 2023 came the *NIST AI Risk Management Framework* (RMF): this framework, to be adopted on a **voluntary** basis, provides detailed guidance for organisations on how to identify, assess, and mitigate **risks** in AI systems, via their mapping, measurement, and management [116]. Today, the NIST AI RMF is for the American tech industry (and its allies) a **baseline** for **governance**

best practices, more so as it aligns with other global standards, such as the *OECD*[1]'s *Principles for Trustworthy AIs* [1](Figure 3.1).

Also in 2023, the White House released an *Executive Order on Safe, Secure, and Trustworthy AI*, effectively the most extensive action by the U.S. government on AI to date[119]. While an Executive Order is not legislation, it still binds the actions of federal agencies in indirect ways, for example, by conditioning federal funding or procurement on certain AI safety practices. This Executive Order in particular shines for establishing a **comprehensive** approach to AI:

- It requires developers of the most powerful AI models to share safety **test results** with the government before deployment[50];
- It charges **NIST** with setting rigorous AI **safety standards** (e.g. for red-team security testing and watermarking AI-generated content)[23];
- It instructs agencies to develop standards to prevent AI from exposing personal data, and to uphold civil **rights** (in decisions on credit, housing, employment, etc.)[12].
- It addresses risks of AI in critical **infrastructure** and bioweapons[50];
- It promotes U.S. **leadership** in setting international AI norms[65].

3.3 The USA's version of a innovation-regulation balance

The United States has long cultivated a historically **liberal**, **market-oriented** philosophy, prioritising support for **innovation over prescriptive regulation**. Executive testimony and academic surveys repeatedly note that U.S. firms view «risk-averse and burdensome regulation»[14] as a direct **threat** to the speed of technological development and innovation-first plans, especially in the IT field. This attitude is reinforced by a **legal environment** that offers strong **property rights** protections, limited **liability** for early-stage ventures, and a tradition of **self-regulation** through industry standards rather than statutory mandates.

Recent efforts to reconcile innovation with oversight have favoured regulatory ***sandboxes*** as experimental spaces where firms can trial new AI services under a temporary, negotiated set of rules. For example, Utah's legal services sandbox, launched in August 2020, is a good example of how sustained dialogue between the regulator and the industry can bring about evidence-based reform that is more naturally accepted by its subjects. However, the limited **scale** of these pilots and the lack of a **coordinated** national framework mean that insights remain fragmented, with the broader U.S. stance continuing to rely on *ad hoc*, light experimentation rather than a unified regulatory architecture.

The U.S. approach favours dynamic, iterative regulation that evolves alongside the industry's actors. Fenwick, Vermeulen and Corrales Compagnucci [48] identify two complementary setups in doing that: near the sandbox-type instruments just mentioned,

[1]Organisation for Economic Co-operation and Development

there is an intense cultivation of innovation **ecosystems** that pair established corporations with startups, allowing the merging of significant innovative pushes with some supervision by established, recognised firms that informally have the power to set the rules in the industry.

Despite the rhetoric of *collaborative* governance, the United States' policy process is heavily shaped by **industry influence**. Lancieri, Edelson and Bechtold [83] document six principal channels through which AI firms **capture the regulatory agenda**, i.e. with their own agenda setting, advocacy, academic capture, information management, cultural capture and media capture, demonstrating a **systematic** ability to steer legislation toward the **weaker** will of government. It is not only the traditional definition of *lobbying* that is enacted: firms also dominate **public-private advisory groups**, draft standards that later become *de facto* regulations, and fund *think tanks* that produce **research** oriented to their preferred policies. The result is a regulatory environment where, in the words of Lévesque[85], «*firms are essentially grading their own homework*,» with «*the concentration of power in a few industry players veer[ing] into self-governance*», at the cost of risking insufficient public interest safeguards. As confirmation, Carvão, Ancheva, Atir et al. [[33]] have examined 150 AI related bills presented before the 118th U.S. Congress and have identified «*emerging areas of alignment*» between policymakers and industry, and have warned that collaboration must be guarded against capture using independent **auditing**, public **reporting** and statutory **enforcement** .

Lévesque's research shows that these safeguards might not be working, at least for now: for example, the author cites the case of the *US Algorithmic Accountability Act*, which requires firms to perform impact assessments; its requirements are so broadly stated that they «*provide practically limitless flexibility [...] in performing impact assessments*», potentially resulting in «***cherry-picking** or manipulation*». This means that firms can, in practice, select methodologies and metrics that cast their actions in the most favourable light, enabled by permissive language that allows for «*subjective, self-serving interpretation*».

To complete the picture, Littler's 2024 *C Suite Survey*[96] of more than 330 senior executives found that 56% of U.S. C suite leaders still do not have an established generative AI **policy** in place. This statistic alone encapsulates the so-called ***implementation gap***, the first manifestation of the the scarce presence of a formal, externally enforced governance. The results of the survey are troubling, especially when considering smaller and mid-sized enterprises, where **resource constraints** and **competing priorities** make formal AI governance seem like a **luxury** rather than a necessity.

Even among the most virtuous firms that have established such policies, severe implementation problems persist: Littler's research identifies that C-level executives are in **disagreement** between themselves on how aggressively the firm should pursue AI, what risks are tolerable, and how compliance should be monitored. As this misalignment cascades downward, mid-level managers face an impossible choice: advance innovation (and immediate profit) or protect the firm from possible harm? In practice, the firm often defaults to the path of least resistance, which is rapid deployment with minimal foresight on collateral risk, as the costs of innovation delay are immediate and visible, while the costs of risk materialisation are, by nature, felt as extremely distant.

3.3.1 Understanding Regulatory Capture[136, 129, 83, 128]

Regulatory capture is a complex phenomenon with **far-reaching implications for public welfare**, particularly in rapidly evolving and wide-impact sectors like Artificial Intelligence.
The concept emerged from George Stigler's [114] seminal work in 1971, in which regulatory capture described situations observed where market incumbents lobbied for stricter rules, such as licensing requirements in publishing, to protect themselves from competition from new entrants. Fundamentally, it occurs when industry actors are able to **«hijack» regulatory regimes** and policymakers, steering them to serve their **private interests** rather than the public good. Today, this encompasses not just the strengthening of regulations to protect particular interests, but also efforts to **weaken oversight**, **under-regulate** harmful practices, or influence how rules are **enforced**.

The role of regulatory capture is especially pronounced in fields such as Artificial Intelligence, where industry actors are on a war path to gain profitability and market dominance, even at the cost of caution. Regulatory capture poses a fundamental threat not just to the average person, but to democratic institutions as well, as a citizen who perceives the law as serving only private interests will lose trust in the government, which in turn will undermine the legitimacy of institutions and will make it even more challenging to implement effective policies that are actually followed. In sum, as priorities become distorted, power becomes entrenched, giving established industry players a permanent advantage and making it harder and harder to promote equality or foster true competition.

In practice, regulatory capture unfolds through a variety of mechanisms, both **overt** and **subtle**.

Direct mechanisms involve immediate and often visible efforts to shape the decisions of policymakers; this is the case of ***personal engagement***, where industry representatives participate in formal policy processes and advocacy, creating the opportunity to stall progress or steer debates in their favour. *Incentive shaping* is another direct method, ranging from campaign **donations** and **gifts** to the notorious *revolving doors* phenomenon, when individuals move between roles in industry and regulatory agencies, retaining contacts and influence wherever they go and exploiting them as needed. *Information capture*, meanwhile, occurs when industry actors manage to manipulate the flow of **information** going towards policymakers, e.g. through agenda-setting [2], selective data sharing, or even sharing overwhelming technical detail. Lastly, lobbying remains significantly effective, with evidence showing great confidence in those efforts, especially in sectors like AI (the scale is substantial: in the United States, AI-related lobbyists increased by 120% in 2023 relative to 2022, with 85% hired by industry).

[2] *Agenda-setting* refers to the ability of actors, in this context AI companies, trade associations, and lobby groups, to shape what policy issues are placed on the decision-makers' radar and which solutions are considered legitimate. It is a popular capture channel [128] that works, for example, by deciding first which risks should be brought up and discussed primarily, and how issues are framed (e.g. by describing a process as an inevitable, uncontrollable force).

Indirect mechanisms of capture are more subtle, shaping the **broader environment** in which regulatory decisions are made. **Academic capture**, for example, involves industry funding of research institutions, subtly influencing academia's hot topics [3]. **Normative capture**, on the other hand, happens when regulators *unconsciously* adopt the values and assumptions of the industry they oversee, such as the belief that innovation is inherently beneficial, as a result of *psycho-social* mechanisms. Something similar is ***cultural capture***, which occurs when regulators are brought into shared networks and social ties to exploit group identity effects and align interests. **Media** and **public relations** close the circle by downplaying risks or amplifying industry-favoured perspectives.

The **effects** of regulatory capture are visible in several aspects of governance, starting from the content of policy itself, which ends up being **weak**, **under-enforced**, or **imbalanced** in prioritising specific goals (often **innovation** or **profit**) at the expense of safety and fairness. In the realm of AI, this translates to the absence of robust safety standards or inadequate oversight for the adverse effects of products. At the **enforcement** stage, even well-crafted policies can be undermined if enforcement is lax, selective, or biased. In its most powerful form, regulatory capture can shape the very structures of governance, resulting in **underfunded supervisory agencies**, fragmented rules, and preemptive federal policies that block what would be more effective state or local action.

It is essential to recognise, however, that **not all industry involvement** in regulation amounts to capture. In fact, industry input can be **beneficial**, providing regulators with valuable technical knowledge and insight on how things work in day-to-day operation in the market they are trying to «tidy up». The line is drawn when intervention ultimately benefits only a few influential players. Experts have suggested a broad range of strategies, such as developing greater technical expertise within policymakers , since meeting the challenge of regulatory capture also requires to reject narratives of governmental incompetence and instead have them embrace their responsibility as stewards of the public good, mustering ambition and confidence to fulfil this delicate role.

[3] For example, the Italian philosopher and professor Luciano Floridi, in his book on the ethics of AI [49] is highly critical of speculative fears about malevolent ultraintelligent machines: he dismisses these as "irresponsibly misleading" distractions that divert attention from the real and urgent ethical challenges posed by current AI technologies. In his view, the true risk is not an imaginary AI monster, but rather human stupidity, malice, and the misuse of powerful but non-intelligent systems.

Chapter 4

China's approach: state-centric, social governance- and sovereignty-focused

China's regulatory experiments are an alternative model of AI governance to that of Western Countries. The strategy is implemented by a series of policies that outline a clear, ambitious **roadmap**, going from foundational **research** to widespread **societal integration** and **global** governance influence, through a dual approach of fostering rapid innovation while establishing robust regulatory **controls** [139].

The paper by Wang et al. titled «*Artificial Intelligence Law(s) in China: Retrospect and Prospect*» [126] describes the country's strategic and adaptive approach to regulation, which prioritises balancing innovation with risk management. Rather than a single, comprehensive AI law, China employs a flexible governance model that combines technical standardisation with sector-specific regulations (e.g. those for autonomous driving or financial markets).

The need to respond rapidly to new issues emerges from China's «*Pilot-First*» regulatory experimentation strategy, involving the use of **pilot** programmes in major tech hubs like Shanghai, Shenzhen, and Beijing, which serve as testing grounds for new AI regulations. In the rest of the country, remains the dual aim of promoting rapid AI development to ensure global competitiveness while mitigating its risks.

4.1 National AI Strategy and Governance

Nowadays, China is entirely on a path to becoming a global AI leader, backed by **strong government directives and a rapidly growing IT industry**, under a tight regulatory regime for AI that prioritises **state control, societal stability, and adherence to Party values**.

China's legislative strategy has been to issue specific regulations on AI technologies, especially those impacting information content and public opinion, and to integrate AI oversight into its already restrictive cyberspace governance framework; in particular, the *Cyberspace Administration of China* plays a key role in content-related AI rules, and the *Ministry of Industry and Information Technology* oversees the industry's development

and standards, also in international settings (e.g. at the ISO and ITU).

Ethical AI guidelines have been promulgated by both government and academic bodies. For example, in 2019, China's *AI Governance Committee* (under the *Ministry of Science and Technology*) released «*Responsible AI for New Generation AI*», a work emphasising **harmony**, **fairness**, and **transparency**. In any case, algorithmic transparency and accountability are pursued primarily in the service of political and social stability (e.g., through the establishment of algorithm registries and content controls) rather than individual rights in the Western sense. The result is a **robust** and **state-centric** regulatory framework: companies must above all ensure their AI is «**trustworthy**» according to government criteria: not amplifying misinformation, not enabling forbidden speech, and not exceeding the carefully drawn borders of government endorsement.

Two other documents form the backbone of China's national AI strategy: the *New Generation Artificial Intelligence Development Plan* [113] (2017) and the *AI Plus Action Plan* [111](2025).

4.1.1 The AIDP

China's *Artificial Intelligence Development Plan* (*AIDP*), unveiled in 2017, sets the long-term vision for China to be the **world leader in AI innovation by 2030**, with an AI industry worth about 150 billion American Dollars [103]. The document frames AI as a **strategic technology** and **a proving ground for international competition** [112]: the following years mark a critical window of opportunity to seize a **first-mover advantage** and accelerate the nation's development into a global science and technology power. Such a comprehensive plan cannot require anything less than a «whole-of-nation» approach, which systematically aligns government, industry, and academia towards that unified goal, through heavy investments guided by four fundamental principles:

- **Research support**: the creation of a world-class AI talent pool through **education** reform and **recruitment** programmes to support long-term research in AI methods, tools, and systems; in particular, the plan emphasises research into areas such as big data **intelligence**, cross-media perception, human-computer **hybrid** intelligence, and autonomous systems. All these have the real possibility of being used in sensitive contexts.

- **Concentration of resources**: the strategy is to take systematic advantage of the socialist system, which can easily concentrate national resources, to successfully gather enough for major AI projects;

- **Market Domination**: accelerate the commercialisation of AI technologies to create a competitive advantage.The government is invested in planning, policy support, security, and ethical regulation;

- **Open Innovation**: sharing between industry and academia is encouraged, as well as collaboration between civilian and military entities. The concept of open innovation has found immense popularity in the West as well, so part of China's efforts in this sense is the participation of Chinese developers and institutions in global, open-source projects, especially in areas of expertise such as natural language processing of Chinese languages.

The 2017 plan also specifically addressed **data governance** as a critical component of its strategy, especially the possibility of making government data more accessible for AI training while still protecting personal data.

4.1.2 The AI Plus Action Plan

The AIDP laid the groundwork for subsequent policies, including the more application-focused «AI Plus» initiative, which lays down the roadmap for integrating Artificial Intelligence **into every facet of the nation's economy and society** over the next decade , according to a phased timeline with ambitious, but quantifiable, targets, starting from the achievement of **deep integration of AI in five key sectors** by 2027 [118]:

- **Science and Technology**: the initiative places strong emphasis on scientific and technological innovation. By equipping researchers with advanced AI tools, the plan aims to **accelerate scientific discovery, upgrade critical infrastructure**, and enhance innovation across all fields. These scientific advances are intended to ripple outward and reach every sphere of human life, fuelling progress in both industry and society.

- **Industrial Development**: The AI Plus initiative is just the latest in a long list of interventions for technology-led industrial modernisation [30] towards the goal of cultivating «new quality productive forces» [34], shifting from a resource-intensive production model to one driven by technological innovation and **high-value-added manufacturing**. The policy's objectives include establishing industry **incubators**, achieving **breakthroughs** in about 100 core technologies [135], and ultimately reduce as much as possible China's **reliance** on foreign powers for high-tech solutions (the reliance on advanced semiconductors is a key vulnerability, due to the push back of strongly regulated exporting countries [9]).
  AI is considered a key driver for China's industrial chain, as it can enhance innovation, productivity, and resilience. Research shows that this path is already delivering its promises: for instance, in the *new energy vehicle* (NEV) sector, the integration of digital twin technology and intelligent decision-making systems has significantly **reduced** R&D cycles and **improved** fault prediction accuracy [60]. Another study [42] on the Chinese Stock Market found that, for an increase in the use of AI of one standard deviation, there has been a corresponding growth of revenue of 5.7% after a year, amounting to roughly 75 million USD.

- **Goods Quality**: new use scenarios open for the advent of intelligent services to boost **productivity** and **convenience**, such as smart assistants that contribute to household management, travel, and elder care. The plan mentions the concept of «*all things intelligent*», which requires the **ubiquitous** presence of smart terminals that take the most disparate forms, like intelligent connected vehicles, AI-powered phones and computers, smart robots, home devices, and wearables, but also low-altitude aircraft and brain-computer interfaces. All this promises to bring a higher **quality of life**.

- **Public Welfare**: the plan seeks to both create **new jobs** and empower **traditional ones**, by strongly supporting AI skills training and favouring re-employment

of those workers who were replaced by innovation.
Society at large is expected to reap AI's benefits, with targeted efforts to improve public services: in **healthcare**, for example, the application of AI is projected to increase diagnostic accuracy and expand access to medical care, particularly in rural areas through **telemedicine** solutions. **Education** stands to benefit as well, with AI enabling large-scale personalised learning and potentially narrowing the educational gap between urban and rural regions, using «intelligent study companions» and «intelligent teachers». AI is supposed to become an integral part not just of technology but of **life and culture** of the Chinese people, even in areas that have been considered, until now, prerogatives of humanity: artistic production, social connection, elder and child care.

- **Governance Capability**: this pillar aims to **modernise public administration and enhance social control** through AI via its integration in municipal infrastructure, urban planning and operations. One of the goals of the CCP is to implement tight monitoring, a public security warning infrastructure, and fast emergency response services. China is actively developing smart cities that utilise AI for the more efficient management of **traffic** and other urban services, set to work in tandem with AI-driven facial recognition and data analysis, which the government frames as essential for maintaining national security and social stability.

4.1.3 Associated Challenges and Risks

Despite its ambitious aims, the AI Plus initiative faces a host of internal and external **challenges**, identified in academic literature, that threaten its successful implementation, ranging from **industrial and social disruption, ethical concerns, and resource constraints**.

Industrial and Social Disruption The plan's success could lead to significant shifts in the employment structure [62]. For example, integration of AI in Chinese workplaces has already been shown to increase workloads and stress for employees, who must meet the demands of both human supervisors and algorithmic metrics [115], as AI solutions are often designed to maximise labour extraction, not to improve working conditions and job quality.

Trustworthiness and Ethical Concerns Widespread AI integration obviously raises the critical issues of data privacy, algorithmic bias, and trustworthiness.

Public perceptions on the matter show distinct characteristics when compared to non-Chinese, particularly Western, approaches (for Brauner et al., [15] these differences are rooted in cultural values, governance philosophies, and public attitudes toward technology):the Chinese public generally displays a high level of **optimism** and **acceptance** towards AI.A survey by the University of Queensland showing that China is a leader in the positive perception of AI [54]:95% of Chinese respondents declared to **accept** the role that AI will have in the future, the highest among countries sampled, with an high percentage expressing optimism about the technology and 67% already willing to **trust** it. Above all, it is cultural models that influence the people's acceptance to be mere

passive users of AI: a comparative study [53] revealed that Chinese respondents considered it more important to engage with AI and less essential to control it, the opposite of what was said by European and American participants.

Chinese academic discourse on AI ethics also presents unique characteristics that reflect popular opinion. Regarding short-term implications, Chinese scholars' concerns largely mirror the content of international ethical guidelines, focusing on issues like justice, privacy, and transparency; however, discussions about long-term implications are hardly comparable as the concept of fairness varies across cultures and requires more than sterile metrics and calculations [59]. Still, there has been a significant growth in quantitative safety research in China as well, and with it, the quality of technical research on frontier AI safety, for example addressing issues like the misuse of AI in biology and chemistry [6].

Resource Constraints The massive **energy consumption** and the high costs of computing power are significant hurdles. China has already been grappling with limited computing resources due to U.S. chip export restrictions; this has forced Beijing to pursue an all-inclusive approach that rallies all national resources, develops resource-efficient AI models, and invests in major infrastructural projects to build its own computing power sources [[5]].

Another obstacle is the shortage of skilled AI talent: although China has thirty AI-focused universities, it reportedly still cannot meet the industry's demand for new hires [134].

Technology-Application Gap A persistent gap remains between the development of advanced AI technology and its practical, widespread application in industry. In the words of a Chinese technology media outlet (Leiphone), there still is an «impassable chasm between technology and implementation in the large model era» [[39]][40], noting that many companies do not actually use the large models that are being developed. Primary obstacles to adoption are the still poor performance of pre-trained models on Chinese language and for specialised industrial tasks [43] [86], but also inadequate infrastructure, particularly in rural areas, which creates a "digital divide" where urban centres benefit from advanced AI while rural regions lag with antiquated systems.**Financial** constraints also pose a significant barrier, especially for smaller public entities and SMEs, which struggle to afford the high costs of AI deployment and specialised expertise the government require of them.

4.2 Regulations on Algorithms and AI Systems

Instead of a single, *omnibus* AI law, Chinese regulators have adopted a "vertical approach" [141], creating specific, binding regulations for different AI applications as they emerge, especially in cases of tools that can have an influence on the opinions and behaviour of the larger public.

Regulations on recommendation algorithms (2022): Effective March 2022, the *Provisions on the Administration of Internet Information Service Algorithmic Recommendations* were jointly issued by several Chinese regulatory bodies, including the *Cyberspace Administration of China* , the *Ministry of Industry and Information Technology*, the *Ministry of Public Security*, and the *State Administration for Market Regulation* [13]. This was one of the world's first sets of rules targeting the algorithms behind **news feeds**, social media **timelines**, and other types of **content suggestion** services. These provisions require algorithm providers to **register** their algorithms with the government, to promote *positive speech* and not endanger national security or social order, and to avoid addictive behaviours or discrimination. Importantly, they give users the right to turn off personalised content recommendations and require algorithmic transparency measures. Companies must also conduct regular audits of their algorithms for misuse.

Interim Measures for Generative AI Services (2023): With the fast and intense surge of generative AI models, China introduced interim rules in mid-2023 to make emergency regulation of the generative AI services available to the larger public, requiring providers to ensure content reflects core socialist values and does not subvert state power or disrupt economic and social order[67]. To control this, the government imposed a **mandatory security assessment** and a filing requirement for generative AI systems deemed to have «*public opinion attributes*» (i.e., capable of influencing the populace), including labels marking AI-generated content and strong safeguards against the generation of inaccurate and prohibited content[100]. Additionally, it was also required to vet training data to avoid illegal content polluting results and intellectual property infringements[29].

Regulations on «Deep Synthesis» (also known as *Deepfakes*) (2023): Effective January 2023, China enacted rules specifically governing deepfake technology and other AI-generated **synthetic media**. These Provisions on the *Administration of Deep Synthesis Internet Services* mandated that any AI-generated or AI-altered content must be clearly **labeled** as such to prevent confusion[26], and providers of the tools used to generate it must **verify users' identities** and ensure the technology is not misused (for fraud, slander, or endangering national security).

A primary driver behind China's deepfake regulations is the perceived threat to **national security**; on top of those already mentioned, concerns include the potential for sensitive training data leaks, the malicious use of AI by hostile state or non-state actors to conduct cyberattacks, and the manipulation of information to achieve political or military objectives [144]. The ability to create convincing but deceptive content is viewed as a significant risk that could lead to public panic, social unrest, and a general erosion of trust in media and government institutions, thereby threatening social and political stability. This aligns with the country's broader need for censorship and information control[55]: by tightly managing technologies that can rapidly create and disseminate convincing falsehoods, the state aims to protect public opinion on itself and prevent the circulation of material deemed harmful.Simultaneously, Chinese law explicitly addresses the harm deepfake technology can inflict on single individuals, framing it as a violation of **personal rights** [58]. In particular, the aim is to counter harms such as the illegal

acquisition and dissemination of private data and the loss to an individual's image and reputation; indeed,these two facets of danger, one public and one private, are not clearly distinct, as the goal of protecting an individual from reputation damage also serves the state's interest in maintaining order and **protecting key individuals** in power, in a context where national interests remain a paramount concern for regulators.

4.3 The importance of regulating first

The *AI Plus* plan is situated within a landscape of **intense geopolitical competition**, particularly against the United States and the European Union. This rivalry unfolds as both a **technological** and a **regulatory race**: China's effort is to shape the emerging global AI order and influence international standards, countering what is known as the «*Brussels Effect*» of EU regulation and the «*California Effect*» of US-led innovation by contributing its own expertise.

4.3.1 The Brussels Effect[3, 14, 107, 132, 44, 56]

In the landscape of global regulatory governance, the European Union has cultivated a powerful mechanism for spreading its influence, known as the *Brussels Effect*. This phenomenon is described [3] as **market-based regulatory exportation**, through which the EU's internal standards and regulations become **global norms**. At its core, the effect is driven by the sheer **scale** and **economic power** of the EU's single market: access to hundreds of millions of consumers is so attractive that multinational corporations often choose to adhere to demanding EU regulations across their entire global operations to avoid the costs of **market segmentation**, or even having to forego the continent entirely.

The effect's diffusion occurs through both *de facto* and *de jure* pathways [14]: the former emerges when extra-European companies comply for **economic efficiency** [132], while the latter occurs **when other jurisdictions** model their own laws on the EU's approach, often due to political or corporate pressures [107]. In both cases,the success of this mechanism hinges on more than just market size, as the EU also possesses significant **regulatory capacity**, i.e. the institutional expertise to create and enforce complex rules [14]. Most notably, European regulation is so exceptionally stringent with respect to its alternatives that compliance with it is sufficient to meet the demands of other jurisdictions without further adjustments [132]. Moreover, its key regulations target inelastic [14] products or manufacturers that cannot easily evade the EU's regulatory reach. Lastly, there is also the condition known as *non-divisibility*, meaning that the products the standards target are difficult or costly to segment by market .

The Brussels effect is just one of several «*effects*» that shape the world's regulations; however, what makes this one stand out is the EU's **consensus-based decision-making process**, which brings **several different stakeholders** together to a binding agreement, a phenomenon that reinforces its credibility and political value in an outsider's eyes [10]. This dimension distinguishes the *Brussels Effect* from the *California Effect*, which lacks a comparable international, synchronous influence and also cannot

stray from US federal law, and the *Beijing Effect*, which works mainly on actors that submit to it out of **necessity** (e.g. to receive much needed financial incentives).

4.3.1.1 Contextualisation within the AI Act

When applied to emerging technologies like Artificial Intelligence, the dynamics at the base of the Brussels Effect are expected to present themselves, but perhaps not quite as expected this time. On the one hand, a strong *de facto* effect is anticipated by academia for several reasons: meeting many of the Act's requirements, especially in the case of General Purpose AI (training data governance, extensive documentation, model evaluations, etc.), requires considerable effort and expenditure, so much so that it is prohibitively expensive for firms to maintain in parallel a separate, non-compliant model for other markets; when it comes to the choice between market differentiation and non-differentiation, the latter is often the most profitable strategy when expanding globally, especially as the remaining markets pose significantly less entry barriers.

On the other hand, the AI Act's ability to trigger a full-fledged Brussels Effect remains contested by some scholars[132], who predict a fragmented «patchwork effect» rather than the establishment of a global standard, citing competition from other «easier» regulatory models like the US's vertical, sector-specific approach or Japan's soft law. Moreover, the AI Act's own **internal complexities**, such as its superimposing classifications, may also hinder its exportability, and the rise of political agendas that value sovereignty and independence may lead nations to drift apart consistently, limiting the seamless translation of EU norms in those contexts.

Among these voices, Martin Ebers [44], Professor of IT Law at the University of Tartu, Estonia, believes that the AI Act is unlikely to trigger a significant *Brussels Effect* because AI regulation is too **complex** and **contested** a case. Unlike the GDPR, which addressed the relatively straightforward matter of privacy, AI encompasses a vast and disparate set of problems, ranging from health and safety to numerous fundamental rights. Furthermore, there is little international consensus on which concrete applications of AI should be regulated (for example, the EU's proposed ban on certain facial recognition technologies contrasts sharply with China's use of them for social stability). This lack of a shared global understanding of the problem prevents the AI Act from serving as a simple, universally applicable model. Also other complexities make the AI Act a tricky legal framework for other countries to adopt: its existence is closely connected to many other EU laws and is meant to work alongside them, as the Act refers to rules from EU product safety laws to classify risks, adds to anti-discrimination laws, and works with data protection rules like the GDPR. If a country tried to use the AI Act without these other essential pieces, Ebers argues, the Act would not be useful. This is the final nail in the coffin for the Act's chances of becoming a global standard *de jure*, like it happened for the GDPR.

The critical perspective of Almada and Radu [3] embrace and goes even further than Ebers', as they warn of a possible Brussels «*Side-Effect*», wherein the exportation of the AI Act's framework could inadvertently **weaken** the global protection of fundamental rights, if those rights are not easily inscribed in the risks recognised by it. As the AI Act was primarily designed as a piece of product safety legislation, it shines at addressing quantifiable dangers to health and safety, mainly in the form of **probabilities**. However, as Almada and Radu point out, the harms that AI can pose to fundamental rights are

often of a completely different nature: they are **systemic** (not easily detected in isolated events), deal most of the damage **cumulatively** to entire social groups **over time**, are not easily **quantifiable** (e.g. damages to dignity, plurality, justice), and are even subject to a society's own **interpretation** of fairness and social relationships.

On the other side of the debate, Greenleaf [56] points how that, for how it is written, EU law contemplates, in a way, its own extra-territorial application, as legislation like the GDPR and the AI Act explicitly impose obligations on non-EU entities whose services or products reach the EU market, dealing with people of any nationality on European soil. Moreover, EU norms have already been incorporated into international agreements and standards, such as those developed by the Council of Europe or the OECD, effectively creating an important precedent for exporting its principles in multilateral forums that see broader participation than the twenty-seven Member States.

4.4 Data Protection

China implemented two major data laws in 2021 that affect AI systems: the *Personal Information Protection Law* (PIPL) and the *Data Security Law* (DSL)[37, 98]. The PIPL establishes comprehensive rules for handling personal information, including the need to obtain consent for use of personal data, minimise data collection, and avoid algorithmic discrimination. Notably, PIPL gives individuals rights not to be subject to automated algorithmic decision-making if the outcome has a significant impact on their rights, and to demand explanations of such algorithmic decisions[7]. Companies using AI for personalised recommendations must also provide options to opt out of targeting[28]. The *Data Security Law* completes the picture by requiring risk assessments for activities handling «*important*» data used in AI training datasets, especially those sourced from the general population.

4.4.1 Regulation of the private sector

China is known for its extensive use of **facial recognition** and other biometrics, especially in public surveillance (e.g. ubiquitous CCTV cameras trained to recognise people of the **Uyghur** ethnic minority in Xinjiang, which has drawn global criticism [38]). There has not been a law banning or comprehensively limiting the use of facial recognition by the government; to the contrary, it remains an invaluable tool . Regulatory and legal developments in recent times all revolve around only the *commercial* use of biometrics with AI intervention [27]. For example, a landmark case in 2021 saw a Chinese court rule in favour of a plaintiff who sued a wildlife park for requiring facial scan entry, invoking consumer rights law and PIPL; after that, cities like Hangzhou introduced guidelines requiring businesses to obtain explicit consent for collecting facial data, proving that even if state use continues broadly, private sector use is being actively curtailed[32].

4.5 AI in Criminal Justice and Adjudication

China has been actively integrating AI into its criminal justice system to assist with **judicial decision-making**[57], an efficiency issue that has been object of reforms since the 1980s. On July 8, 2017, the State Council's *New Generation Artificial Intelligence Development Plan* explicitly called for the establishment of «*Smart Courts*», a concept first proposed in 2016 by Zhou Qiang, the president of the *Supreme People's Court*, and the promotion of «*AI applications for applications including evidence collection, case analysis, and legal document reading and analysis*» [127]. Their implementation in criminal proceedings, an area that traditionally requires great caution with technological innovation, highlights the aggressive approach Chinese courts are taking toward adopting AI technologies, a part of a renewed wave of nationwide effort. On their end, researchers and policymakers are focused on establishing ethical norms for judicial AI, clarifying its role as an **auxiliary** tool, and defining the **scope** of its application to mitigate risks and defects. The overall aim is to balance judicial fairness with technological advancement within a robust regulatory framework. A prominent example of this is the Shanghai AI-Assisted Criminal Case Handling System, also known as the «206» system [133]. This system, directly supported by China's *Central Political and Legal Affairs Commission*, is a data-driven, machine learning-based tool that assists legal professionals by identifying **commonalities** among cases to reportedly enhance **consistency** and **accuracy** in handling new ones. The plan sees these tools being used in tasks including evidence collection, case analysis, reading legal documents, assisting judges in sentencing, and automatically generating documents.

Chapter 5

Russia

Russia has identified the development of Artificial Intelligence as a key **state policy priority**, and a critical factor in national **prestige**, economic **competitiveness**, and **military strength** on the global stage [91]. The country's approach is characterised by a strong ideological drive to secure its position as a global power and a pragmatic, evolving legislative framework designed to foster growth while managing **risks to internal social stability**. Without doubt, the Russian leadership's interest in AI is deeply tied to its **geopolitical ambitions**: President Vladimir Putin himself stated that «*Whoever becomes the leader in this sphere will become the ruler of the world*»[123], a sentiment that frames the global AI race as a competition for much more than market share. In fact, Russian authorities have gone on the record saying that technological development, including AI and unmanned equipment (especially **combat robots**), is essential to preserve the *Russian civilisation* [143] **at an existential level** as AI is a critical component of national security strategy and a key factor in maintaining Russia's position on the global stage together with a combination of strategic planning, surveillance of perceived threats [142], and policies for military modernisation. The 2019 adoption of a national AI strategy further solidified this commitment, creating a guiding framework for accelerating AI applications nationwide [101].

5.1 Military Strategic Stance on AI

Russia's stance on AI is made much more interesting due to its status as a country at war, both ideologically and physically: the **Russo-Ukrainian War**, ongoing since 2014, has seen many instances in which autonomous devices and «*agents*» have been used to obliterate human life, and its long standing tug of war with Western democracies, especially the United States of America, still evolves daily in the form of deepfakes and the proliferation of human-like bots on social media, mainly as propaganda and as efforts to divide and inflame the political debate. As far as combat applications are concerned, the goal is to develop relatively inexpensive but capable forces to address the current need to **replace human soldiers** in dangerous situations[90].

Russia's approach to AI development reflects the strategic importance the Kremlin places on AI as a cornerstone of national security and prosperity: it is predominantly state-driven and top-down, with key institutions involved or even established *ex-novo* to lead this effort, including

The Advanced Research Foundation (ARF), tasked with developing the next generation of weapons, and the Ministry of Defence AI Department. The entire effort, from funding and research to implementation, is centrally organised, backed and directed personally[82] from the highest levels of the Russian government, and has now been the object of high-level strategic plan-making for almost ten years now [1].

The "human-in-the-loop" principle Within the Russian military establishment, there is an ongoing discussion about the desired level of **autonomy** for AI systems[92]: some Russian policymakers and military figures advocate for embracing a trend toward automation and robotisation for the reduction of human involvement in weapons systems control and data analysis, while others have expressed a more cautious approach, resulting in Russia's official position at the United Nations on the «*[...]necessity of maintaining* ***human control*** *over the machine*»[47].

However, declarations of Russian military vertices imply they see total autonomy is in the cards due to their unparalleled speed and accuracy with respect to human beings [31], making real people a liability in «machine-paced» situations . The eventual dismissal of the *human-in-the-loop* principle is even more probable if we recall that Russia sees technology as the ticket to supremacy in the geo-political world: since other countries, as is the case with the Turkish *Kargu 2* drones in Libya [31], are upgrading in the direction of total autonomy, escalation from the Russian military to retain competitiveness would be only natural .

5.1.1 The Propaganda machine

Russian state-affiliated actors are increasingly integrating Artificial Intelligence into their disinformation and propaganda efforts [125]. The development of AI has provided new, powerful tools for information warfare, allowing **centralised groups** funded directly by the Russian government to disseminate manipulative content at an unprecedented scale; in particular, the Russo-Ukrainian war has served as a prime example of AI's use in wartime propaganda, especially «*deepfakes*» that manipulate facts and spread disinformation [94, 104] on political leaders to undermine their credibility; for instance, deepfakes have been used to depict Ukrainian leaders making false statements and to deceive high-ranking international officials, even in real time (e.g. via video calling) [109]. AI is also used to generate professional-sounding voiceovers for videos to enhance their perceived credibility; tools like ChatGPT have been employed to mass-produce fake **news articles**, social media posts, and online comments, as in the ***CopyCop*** campaign, which used AI to scrape articles from legitimate news websites, rewrite them with a pro-Trump and right-wing **bias**, and publish them on a network of fake news sites staffed by AI-invented journalists[11]. The primary goals of these campaigns are to manipulate public opinion, sow distrust, destabilise adversaries, and justify Russia's geopolitical actions [11, 4]. Key targets include:

- The War in Ukraine: the goal is to reduce international support for Ukraine, discredit its leadership, and promote pro-Kremlin narratives about the conflict.

[1] In December 2016, the government adopted the *Strategy of Scientific and Technological Development of the Russian Federation*, which explicitly mentions Artificial Intelligence.

- Western Elections: campaigns seek to influence election outcomes in the United States and Europe by exacerbating existing political divisions, promoting favoured candidates, and spreading rumours of election fraud perpetrated by undesired candidates.

In particular, several large-scale, Russian-linked influence operations have been identified as heavily reliant on AI:

- *Doppelganger*: a vast network of **fake news websites** and **social media bots** designed to disseminate Kremlin-friendly narratives. The network has been linked to sanctioned Russian IT firms operating at the Kremlin's direction. Doppelganger has targeted European and American audiences with disinformation about aid to Ukraine, the Paris Olympics (including an AI-generated voice of actor Tom Cruise narrating a fake documentary disparaging the *International Olympic Committee* [93]), and US politics, with videos mocking President Joe Biden. [79, 2]

- *Networks Operation Overload* (also known as *Matryoshka* or *Storm-1679*): This campaign, active since 2023, impersonates trusted media outlets and academics to lend credibility to its narratives. It is supported by AI-generated voiceovers and includes fake leads for fact-checkers, to infiltrate their sources [87].

- *DC Weekly*[137]: A now-exposed fake news website that purported to be a US-based outlet. The site utilised AI to rewrite content from other sources, enabling it to more than double its article output with respect to a legitimate website and diversify its topics, thereby making it appear more like an official news source. It successfully shared false narratives about Ukrainian corruption that were shared tens of thousands of times, including by members of the US Congress[35].

Lifelike contents generated with AI very effectively blur the line between fact and fiction: even in cases where the result is unrefined and, to expert eyes, clearly fabricated, research [97, 22] shows it can still plant a **subconscious** idea, create long-persisting [106] **false memories**, and influence beliefs and decisions [36] with practical and very believable «*pseudo-explanations*» that give the illusion of logical reasoning. Furthermore, the very existence of deepfake technology supports the *liar's dividend* [24] effect, where public figures can dismiss genuine content simply by affirming that they are actually very well-made fakes.

While their effect can vary primarily based on individual predisposition[140], campaigns that blend authentic media with AI-generated enhancements have proven more successful than «traditional» ones, making fabricated propaganda appear more legitimate [124] and posing a real threat to global information integrity.

Apart from the generation of lifelike videos and pictures, AI can also significantly **enhance existing techniques**, for example by employing bot farms to create fictitious online personas that spread pro-Kremlin messages across social media, to create a false impression of popular support for an event and inflame public debate on social media platforms [80]. Thanks to AI, their language and discourse can adapt to a specific post or video without human intervention, making them essentially indistinguishable from real users [41].

AI models as well can be the targets of propaganda attacks: an extremely covert and dangerous type of attack is *poisoning* AI Models (also called LLM *Grooming*) by flooding the Internet, which is continuously scraped for new training information and real-time grounding. There are networks like *Pravda* with the specific goal of making AI inadvertently output Russian propaganda when users ask chatbots questions [130][99]. Studies have already confirmed that extremely powerful chatbots are indeed vulnerable to this attack, as they sometimes base their responses on Russian-linked fake news sites [102].

5.2 Internal control

At the same time as these propagandist efforts, Russian officials continue to express concern about Western dominance in AI development, fearing it could be used to subvert Russian public opinion and destabilise the domestic information environment [138]. The state's tight control on AI is driven by ideological imperatives that identify Western AI models as a threat that could compete with Russian culture and champion ideas that the Kremlin opposes; this risk has fuelled a push for self-reliance and a *sovereign Russian Artificial Intelligence*[81], **developed and deployed by the state**, or by collaborators in the private sector that align with current political and ideological goals.

Chapter 6

Canada's Public Sector Algorithmic Transparency

Canada has emerged as a pioneer in **addressing AI in government operations**[8]. Its legislative framework, following the primary strategy of creating large, non-sector-specific regulations on AI itself rather than a specific application, centres on the *Directive on Automated Decision-Making* (DADM) dealing with the government's employment of AI and the use of the *Algorithmic Impact Assessment* (AIA) tool for harm mitigation[1]. For other sectors, the frequent recourse to soft law emphasises a preference for guidance on delicate themes over strict enforcement, and a focus on concrete applications.

The DADM directive, much like the EU's AI Act, employs **a risk-based approach**, requirements tailored to the specific risks associated with the use of AI technology. Among these, a central *mechanism* to mitigate potential harms emerges clearly [108]: **transparency**. For Canada to keep the accountability and legitimacy of democratic processes that the Country is automating, transparency is applied in six different ways:

1. The possibility for meaningful human oversight and intervention,
2. Mandatory *Algorithmic Impact Assessments*,
3. Regular auditing to evaluate compliance,
4. Disclosure to stakeholders,
5. Inventories of AI systems in use,
6. and adversarial testing.

Transparency is seen as a crucial ally for regulators and lawmakers to be able to effectively oversee operations, understand them, and eventually identify problems, because when individuals understand how AI systems affect them, they can more effectively challenge their decisions and seek **redress**.

Apart from legislative efforts, Canada has implemented a significant number of initiatives to fund **research** and support AI adoption over the long term; the lion's share of

[1] A questionnaire for governmental officials to assess the impact level of an automated decision-making system before it is put into the real world.

them is definitely dedicated to innovation, support for startups and infrastructure building and productivity, while domains such as worker reskilling, education and ethics remain, in practice, underrepresented[8]. Canada's approach is, therefore, still innovation-first (but not, interestingly, *standard-setting-first* too), with a focus on building economic **competitiveness**, both in the private and the public sector, with few but clearly defined ethical safeguards in place (there are regional variations to this, depending on the single province's industrial vocation).In short, Canada's approach to public sector algorithmic transparency presents a paradox: on one hand, the nation stands as a global pioneer; on the other, this progressive architecture is built on a **soft foundation**, resulting in a system that excels in articulating its moral principles but falls short in strongly guiding their practical implementation.

Bibliography

[1] *AI Principles Overview — oecd.ai*. https://oecd.ai/en/ai-principles. [Accessed 18-11-2025].

[2] Berkant Akkuş. "Deepfakes and the Geneva Conventions: Does Deceptive AI-Generated Misinformation Directed at an Enemy During Armed Conflict Violate International Humanitarian Law? A Critical Discussion". In: *Laws* 14.6 (Nov. 2025), p. 83. ISSN: 2075-471X. DOI: 10.3390/laws14060083. URL: http://dx.doi.org/10.3390/laws14060083.

[3] Marco Almada and Anca Radu. "The Brussels Side-Effect: How the AI Act Can Reduce the Global Reach of EU Policy". In: *German Law Journal* 25.4 (Feb. 2024), pp. 646–663. ISSN: 2071-8322. DOI: 10.1017/glj.2023.108. URL: http://dx.doi.org/10.1017/glj.2023.108.

[4] Maxim Alyukov, Maria Kunilovskaya and Andrei Semenov. "Wartime Media Monitor (WarMM-2022): A Study of Information Manipulation on Russian Social Media during the Russia-Ukraine War". In: *Proceedings of the 7th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Association for Computational Linguistics, 2023. DOI: 10.18653/v1/2023.latechclfl-1.17. URL: http://dx.doi.org/10.18653/v1/2023.latechclfl-1.17.

[5] Anonymous. *China's AI Development Model in an Era of Technological Deglobalization*. Report. Mercator Institute for China Studies (MERICS). URL: https://merics.org/en/report/chinas-ai-development-model-era-technological-deglobalization (visited on 18/11/2025).

[6] Anonymous. *New Spring 2024 updated report on State of AI Safety in China*. 15th May 2024. URL: https://aisafetychina.substack.com/p/new-spring-2024-updated-report-on (visited on 18/11/2025).

[7] *Article 24 – Automated Decision Making*. Accessed: 2025-11-18. 2021. URL: https://pipl.xllawconsulting.com/personal-information-protection-law-of-the-peoples-republic-of-china-pipl/chapter-ii-personal-information-processing-rules/section-1-general-rules/article-24/.

[8] Blair Attard-Frost, Ana Brandusescu and Kelly Lyons. "The Governance of Artificial Intelligence in Canada: Findings and Opportunities from a Review of 84 AI Governance Initiatives". In: *SSRN Electronic Journal* (2023). ISSN: 1556-5068. DOI: 10.2139/ssrn.4414212. URL: http://dx.doi.org/10.2139/ssrn.4414212.

[9] Placeholder Author(s). *Artificial Intelligence in China: A Multidisciplinary Comprehensive Analysis of Origins, Development, Actors, Strategies, Ambitions, Capabilities, and its Intersection with Semiconductors*. Please fill in the actual author(s) and publication year from the webpage. HERMES-Kalamos. Placeholder Year. URL: https://www.hermes-kalamos.eu/artificial-intelligence-in-china-a-multidisciplinary-comprehensive-analysis-of-origins-development-actors-strategies-ambitions-capabilities-and-its-intersection-with-semiconductors/ (visited on 18/11/2025).

[10] Annegret Bendiek and Isabella Stuerzer. "The Brussels Effect, European Regulatory Power and Political Capital: Evidence for Mutually Reinforcing Internal and External Dimensions of the Brussels Effect from the European Digital Policy Debate". In: *Digital Society* 2.1 (Jan. 2023). ISSN: 2731-4669. DOI: 10.1007/s44206-022-00031-1. URL: http://dx.doi.org/10.1007/s44206-022-00031-1.

[11] Vera Bergengruen. *OpenAI Says Russia, China, and Israel Are Using Its Tools for Foreign Influence Campaigns*. Time. 30th May 2024. URL: https://time.com/6983903/openai-foreign-influence-campaigns-artificial-intelligence/ (visited on 18/11/2025).

[12] Jr. Biden Joseph R. *Executive Order 14110: Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence*. https://www.presidency.ucsb.edu/documents/executive-order-14110-safe-secure-and-trustworthy-development-and-use-artificial. Published in Federal Register Nov 1, 2023. 30th Oct. 2023.

[13] Bird and Bird. *China: Data and evolving digital regulation: algorithm regulation*. Nov. 2023. URL: https://www.twobirds.com/en/capabilities/practices/digital-rights-and-assets/apac-dra/apac-dsd/data-as-a-key-digital-asset/china/data-and-evolving-digital-regulation-algorithm-regulation.

[14] Anu Bradford. "The False Choice Between Digital Regulation and Innovation". In: *SSRN Electronic Journal* (2024). ISSN: 1556-5068. DOI: 10.2139/ssrn.4753107. URL: http://dx.doi.org/10.2139/ssrn.4753107.

[15] Philipp Brauner et al. *Cultural Dimensions of AI Perception: Charting Expectations, Risks, Benefits, Tradeoffs, and Value in Germany and China*. 2024. eprint: arXiv:2412.13841.

[16] Dan Breznitz and Michael Murphree. "Standardized Confusion? The Political Logic of China's Technology Standards Policy". In: *SSRN Electronic Journal* (2011). ISSN: 1556-5068. DOI: 10.2139/ssrn.1767082. URL: http://dx.doi.org/10.2139/ssrn.1767082.

[17] Dan Breznitz and Michael Murphree. *The Rise of China in Technology Standards: New Norms in Old Institutions*. Tech. rep. Accessed: 2025-10-06. U.S.-China Economic and Security Review Commission, 16th Jan. 2013. URL: https://www.uscc.gov/sites/default/files/Research/RiseofChinainTechnologyStandards.pdf.

[18] Evgeniy Bryndin Russia. "Standardization of Artificial Intelligence for the Development and Use of Intelligent Systems". In: *Advances in Wireless Communications and Networks* 6.1 (2020), p. 1. ISSN: 2575-5951. DOI: 10.11648/j.awcn.20200601.11. URL: http://dx.doi.org/10.11648/j.awcn.20200601.11.

[19] California Department of Justice – Office of the Attorney General. *California Consumer Privacy Act (CCPA)*. Accessed: 2025-11-19. 2020. URL: https://oag.ca.gov/privacy/ccpa.

[20] California Privacy Protection Agency. *A New Landmark for Consumer Control Over Their Personal Information: CPPA Proposes Regulatory Framework for Automated Decision-making Technology*. Accessed: 2025-11-19. 2023. URL: https://cppa.ca.gov/announcements/2023/20231127.html.

[21] Andrea Joy Campbell. *AG Campbell Issues Advisory on How State Consumer Protection and Other Laws Apply to Artificial Intelligence*. Office of the Attorney General of Massachusetts. 16th Apr. 2024. URL: https://www.mass.gov/news/ag-campbell-issues-advisory-providing-guidance-on-how-state-consumer-protection-and-other-laws-apply-to-artificial-intelligence.

[22] Samantha Chan et al. *Conversational AI Powered by Large Language Models Amplifies False Memories in Witness Interviews*. 2024. eprint: arXiv:2408.04681.

[23] Bilva Chandra, Jesse Dunietz and Kathleen Roberts. *Reducing Risks Posed by Synthetic Content: An Overview of Technical Approaches to Digital Content Transparency*. National Institute of Standards and Technology, 20th Nov. 2024. DOI: 10.6028/NIST.AI.100-4. URL: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-4.pdf.

[24] Robert Chesney and Danielle Keats Citron. "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security". In: *SSRN Electronic Journal* (2018). ISSN: 1556-5068. DOI: 10.2139/ssrn.3213954. URL: http://dx.doi.org/10.2139/ssrn.3213954.

[25] Rohan Chhatre and Seema Singh. "Policy and Regulatory Frameworks for Artificial Intelligence". In: *SSRN Electronic Journal* (2024). ISSN: 1556-5068. DOI: 10.2139/ssrn.4848705. URL: http://dx.doi.org/10.2139/ssrn.4848705.

[26] *China is about to get tougher on deepfakes in an unprecedented way. Here's what the rules mean*. CNBC, 23December2022. Accessed: 2025-11-18. 2022. URL: https://www.cnbc.com/2022/12/23/china-is-bringing-in-first-of-its-kind-regulation-on-deepfakes.html.

[27] *China Limits Private Use of Facial Recognition on 'Individuals Who Do Not Agree,' but Exempts Security Agencies*. Accessed: 2025-11-18. 2025. URL: https://www.eweek.com/news/china-cyberspace-administration-facial-recognition-ruling/.

[28] *China Releases Draft Amendments to the Personal Information Protection Standard*. Accessed: 2025-11-18. 2021. URL: https://www.insideprivacy.com/international/china/china-releases-draft-amendments-to-the-personal-information-protection-standard/.

[29] *China: Generative AI Measures Finalized*. Accessed: 2025-11-18. Library of Congress. 2023. URL: https://www.loc.gov/item/global-legal-monitor/2023-07-18/china-generative-ai-measures-finalized/.

[30] Changyong Choi and Jihye Yoon. "AI policy in action: the Chinese experience in global perspective". In: *Journal of Policy Studies* 40.2 (June 2025), pp. 1–23. ISSN: 2800-0714. DOI: 10.52372/jps.e685. URL: http://dx.doi.org/10.52372/jps.e685.

[31] CNA. *AI and Autonomy in Russia, Special Issue*. Tech. rep. CNA, Sept. 2022. URL: https://www.cna.org/Newsletters/Ai%20and%20Autonomy%20in%20Russia/AI-and-Autonomy-in-Russia-Special-Issue-September-2022.pdf (visited on 29/09/2025).

[32] *Court upholds ruling in facial recognition case*. Accessed: 2025-11-18. 2021. URL: https://govt.chinadaily.com.cn/s/202104/12/WS6073b782498e7a02c6f6aa95/court-upholds-ruling-in-facial-recognition-case.html.

[33] Allan Dafoe, Jonathan G. F. Scharowski and Carina Prunkl. *The AI Arms Race: The Strategic Dynamics of Artificial Intelligence*. Working Paper AWP-251-2. Harvard Kennedy School, Belfer Center for Science and International Affairs, Mar. 2021. URL: https://www.hks.harvard.edu/sites/default/files/Final_AWP_251_2.pdf (visited on 18/11/2025).

[34] Doreen Dai et al. "Innovation Ecosystems, New Quality in China, AI+, and Corporate Universities". In: *Proceedings of the Americas Conference on Information Systems (AMCIS) 2024 TREOs*. AMCIS 2024 TREOs. Salt Lake City, Utah: Association for Information Systems (AIS), 2024, p. 166. URL: https://aisel.aisnet.org/treos_amcis2024/166 (visited on 18/11/2025).

[35] W. R. Dailey, S. T. Eady and M. R. T. S. T. Linder. "Generative propaganda: Evidence of AI's impact from a state-backed disinformation campaign". In: *Proceedings of the National Academy of Sciences Nexus* 4 (4 2024), pgaf083. URL: https://academic.oup.com/pnasnexus/article/4/4/pgaf083/8097936 (visited on 18/11/2025).

[36] Valdemar Danry et al. "Deceptive Explanations by Large Language Models Lead People to Change their Beliefs About Misinformation More Often than Honest Explanations". In: *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems*. CHI '25. ACM, Apr. 2025, pp. 1–31. DOI: 10.1145/3706598.3713408. URL: http://dx.doi.org/10.1145/3706598.3713408.

[37] *Data Security Law of the People's Republic of China*. Accessed: 2025-11-18. 2021. URL: https://en.wikipedia.org/wiki/Data_Security_Law_of_the_People%27s_Republic_of_China.

[38] *Digital Technologies and Atrocity Crimes in China*. Accessed: 2025-11-18. Global Centre for the Responsibility to Protect. 2024. URL: https://www.globalr2p.org/wp-content/uploads/2024/03/2024-March-Digital-Technologies-Policy-Brief.pdf.

[39] Jeffrey Ding. *China AI 236: The LLM Implementation Gap*. https://chinai.substack.com/p/chinai-236-the-llm-implementation. [Accessed 18-11-2025].

[40] Jeffrey Ding. *China's AI Implementation Gap China Leadership Monitor*. https://www.prcleader.org/post/china-s-ai-implementation-gap. [Accessed 18-11-2025].

[41] Volodymyr Donets. *AI-Generated Content Indistinguishable from Human-Created*. donets.org. 5th Aug. 2024. URL: https://donets.org/risks/ai-generated-content-indistinguishable-from-human-created (visited on 18/11/2025).

[42] Shuyi Dong, Wang Jin and Tianshu Sun. *AI Investment and Firm Performance: Insights from China*. SSRN Scholarly Paper 5055518. Available at SSRN: 5055518, 14th Dec. 2024. URL: https://ssrn.com/abstract=5055518 (visited on 18/11/2025).

[43] Xinrun Du et al. *Chinese Tiny LLM: Pretraining a Chinese-Centric Large Language Model*. 2024. DOI: 10.48550/ARXIV.2404.04167. URL: https://arxiv.org/abs/2404.04167.

[44] Martin Ebers. "Truly Risk-based Regulation of Artificial Intelligence How to Implement the EU's AI Act". In: *European Journal of Risk Regulation* 16.2 (Nov. 2024), pp. 684–703. ISSN: 2190-8249. DOI: 10.1017/err.2024.78. URL: http://dx.doi.org/10.1017/err.2024.78.

[45] The Economist. "China is writing the world's technology rules". In: *The Economist* (Oct. 2024). Accessed: 2025-10-06. URL: https://www.economist.com/business/2024/10/10/china-is-writing-the-worlds-technology-rules.

[46] EncodeAI. *First-Tier Companies Make Standards — Catching Up with China on the Edge*. https://encodeai.org/first-tier-companies-make-standards-catching-up-with-china-on-the-edge/. Accessed: 2025-10-06. n.d.

[47] Russian Federation. *Russia's Approaches to the Elaboration of a Working Definition and Basic Functions of Lethal Autonomous Weapon Systems in the Context of the Purposes and Objectives of the Convention*. Working paper CCW/GGE.1/2018/WP.6 submitted to the Group of Governmental Experts on Lethal Autonomous Weapons Systems, Convention on Certain Conventional Weapons. Apr. 2018. URL: https://unoda-documents-library.s3.amazonaws.com/Convention_on_Certain_Conventional_Weapons_-_Group_of_Governmental_Experts_(2018)/CCW_GGE.1_2018_WP.6_E.pdf.

[48] Mark Fenwick, Erik P. M. Vermeulen and Marcelo Corrales Compagnucci. *Business and Regulatory Responses to Artificial Intelligence: Dynamic Regulation, Innovation Ecosystems and the Strategic Management of Disruptive Technology*. 2024. DOI: 10.48550/ARXIV.2407.19439. URL: https://arxiv.org/abs/2407.19439.

[49] Luciano Floridi. *The ethics of artificial intelligence*. en. London, England: Oxford University Press, Aug. 2023.

[50] Vanessa Friedman. *Biden Hails 'Bold Action' of U.S. Government with Order on Safe Use of AI*. 30th Oct. 2023. URL: https://www.theguardian.com/technology/2023/oct/30/biden-orders-tech-firms-to-share-ai-safety-test-results-with-us-government.

[51] *FTC Finalizes Order Prohibiting IntelliVision from Making Deceptive Claims About Its Facial Recognition Software*. Federal Trade Commission. 13th Jan. 2025. URL: https://www.ftc.gov/news-events/news/press-releases/2025/01/ftc-finalizes-order-prohibiting-intellivision-making-deceptive-claims-about-its-facial-recognition.

[52] Chlotia Garrison and Clovia Hamilton. "A comparative analysis of the EU GDPR to the US's breach notifications". In: *Information and Communications Technology Law* 28.1 (Jan. 2019), pp. 99–114. ISSN: 1469-8404. DOI: 10.1080/13600834.2019.1571473. URL: http://dx.doi.org/10.1080/13600834.2019.1571473.

[53] Xiao Ge et al. "How Culture Shapes What People Want From AI". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. CHI '24. ACM, May 2024, pp. 1–15. DOI: 10.1145/3613904.3642660. URL: http://dx.doi.org/10.1145/3613904.3642660.

[54] Nicole Gillespie et al. *Trust in Artificial Intelligence: A global study*. Feb. 2023. DOI: 10.14264/00d3c94. URL: http://dx.doi.org/10.14264/00d3c94.

[55] Kieran Green et al. *Censorship Practices of the People's Republic of China*. Tech. rep. U.S.-China Economic and Security Review Commission, Feb. 2024. URL: https://www.uscc.gov/sites/default/files/2024-02/Censorship_Practices_of_the_Peoples_Republic_of_China.pdf (visited on 29/09/2025).

[56] Graham Greenleaf. "EU AI Act: Brussels Effect(s) or a Race to the Bottom?" In: *Privacy Laws and Business International Report* (1st Jan. 2024). 10 pages, pp. 1–10.

[57] Zhiyuan Guo and Jiajia Yang. "The Application of Artificial Intelligence in China's Criminal Justice System". In: *Legal Issues in the Digital Age* 6.1 (May 2025), pp. 83–104. DOI: 10.17323/2713-2749.2025.1.83.104. URL: https://lida.hse.ru/article/view/26904.

[58] Mengqi Han. "The Infringement of Deepfake Technology on Personal Privacy and Legal Protection: A Discussion Based on Article 1032 of the Civil Code". In: *Journal of Education, Humanities and Social Sciences* 41 (Oct. 2024), pp. 188–197. DOI: 10.54097/s0a47e08. URL: https://drpress.org/ojs/index.php/EHSS/article/view/26776.

[59] Xin Han, Marten H. L. Kaas and Cuizhu Dawn Wang. "A Cross-Cultural Examination of Fairness Beliefs in Human-AI Interaction". In: (2025). DOI: 10.2139/ssrn.5116823. URL: http://dx.doi.org/10.2139/ssrn.5116823.

[60] Qiujie He et al. "The impact of artificial intelligence on industrial chain resilience: evidence from China's manufacturing industry". In: (Apr. 2025). DOI: 10.21203/rs.3.rs-6253455/v1. URL: http://dx.doi.org/10.21203/rs.3.rs-6253455/v1.

[61] Witold Henisz. "The Political Economy of Trans-Pacific Business Linkages". In: *Business and Politics* 6.1 (Apr. 2004), pp. 1–35. ISSN: 1469-3569. DOI: 10.2202/1469-3569.1083. URL: http://dx.doi.org/10.2202/1469-3569.1083.

[62] Qingqing Huo, Jing Ruan and Yan Cui. ""Machine replacement" or "job creation": How does artificial intelligence impact employment patterns in China's manufacturing industry?" In: *Frontiers in Artificial Intelligence* 7 (Mar. 2024). ISSN: 2624-8212. DOI: 10.3389/frai.2024.1337264. URL: http://dx.doi.org/10.3389/frai.2024.1337264.

[63] IAPP. *Global AI Law and Policy Tracker*. https://iapp.org/resources/article/global-ai-legislation-tracker/. Accessed: 2025-09-20.

[64] "IEEE Standard for Performance Benchmarking for Artificial Intelligence Server Systems". In: *IEEE Std 2937-2022* (2022), pp. 1–54. DOI: 10.1109/IEEESTD.2022.9930948.

[65] ASME Government Relations & Policy Impact. *A.I. Report Highlights Importance of U.S. Standards Leadership*. 2023. URL: https://www.asme.org/government-relations/policy-impact/a-i-report-highlights-importance-of-u-s-standards-leadership.

[66] Axis Intelligence. *AI Standards Guide 2025*. https://axis-intelligence.com/ai-standards-guide-2025/. Accessed: 2025-10-06. n.d.

[67] *Interim Measures for the Management of Generative Artificial Intelligence Services*. Accessed: 2025-11-18. 2023. URL: https://www.loc.gov/item/global-legal-monitor/2023-07-18/china-generative-ai-measures-finalized/.

[68] International Organization for Standardization and International Electrotechnical Commission. *ISO/IEC 23894:2023 – Artificial intelligence — Guidance on risk management*. Geneva, Switzerland, 2023. URL: https://www.iso.org/standard/77304.html.

[69] International Organization for Standardization and International Electrotechnical Commission. *ISO/IEC 42001:2023 – Artificial intelligence — Management system*. Geneva, Switzerland, 2023. URL: https://www.iso.org/standard/81230.html.

[70] International Organization for Standardization and International Electrotechnical Commission. *ISO/IEC TR 20226:2025 – Information technology — Artificial intelligence — Environmental sustainability aspects of AI systems*. Technical Report. Accessed: 2025-10-06. Geneva, Switzerland, 2025. URL: https://www.iso.org/standard/86177.html.

[71] International Organization for Standardization and International Electrotechnical Commission. *ISO/IEC TR 24027:2021 – Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making*. Technical Report. Geneva, Switzerland, 2021. URL: https://www.iso.org/standard/77607.html.

[72] International Organization for Standardization and International Electrotechnical Commission. *ISO/IEC TR 24028:2020 – Information technology — Artificial intelligence — Overview of trustworthiness in artificial intelligence*. Technical Report. Accessed: 2025-10-06. Geneva, Switzerland, 2020. URL: https://www.iso.org/standard/77608.html.

[73] International Organization for Standardization and International Electrotechnical Commission. *ISO/IEC TR 24368:2022 – Information technology — Artificial intelligence — Overview of ethical and societal concerns*. Technical Report. Accessed: 2025-10-06. Geneva, Switzerland, 2022. URL: https://www.iso.org/standard/78507.html.

[74] International Telecommunication Union (ITU). *Artificial Intelligence – ITU Action*. https://www.itu.int/en/action/ai/Pages/default.aspx. Accessed: 2025-10-06. n.d.

[75] International Telecommunication Union (ITU). *Detecting Deepfakes and Generative AI: Report on Standards for AI Watermarking and Multimedia Authenticity Workshop*. Report. Accessed: 2025-10-06. Geneva, Switzerland, 2024. URL: https://www.itu.int/dms_pub/itu-t/opb/ai4g/T-AI4G-AI4GOOD-2024-7-PDF-E.pdf.

[76] International Telecommunication Union (ITU). *ITU-T Work Programme (ISN 15175) — Machine learning in future networks including IMT-2020: use cases*. https://www.itu.int/itu-t/workprog/wp_item.aspx?isn=15175. Last updated: 2022-01-21; Accessed: 2025-10-06. 2019.

[77] International Telecommunication Union (ITU). *Recommendation ITU-T H.870 – Artificial intelligence (AI) — Overview of trustworthiness in AI systems (03/2022)*. Recommendation. Accessed: 2025-10-06. Geneva, Switzerland, Mar. 2022. URL: https://www.itu.int/rec/T-REC-H.870-202203-I/en.

[78] International Telecommunication Union (ITU). *Recommendations Y-series (Y.-series) — Future Networks / NGN and AI-related aspects*. https://www.itu.int/rec/T-REC-Y/en. Accessed: 2025-10-06. n.d.

[79] *Justice Department Disrupts Covert Russian Government-Sponsored Foreign Malign Influence Operation Targeting Audiences in the United States and Elsewhere*. U.S. Department of Justice. 14th Feb. 2024. URL: https://www.justice.gov/archives/opa/pr/justice-department-disrupts-covert-russian-government-sponsored-foreign-malign-influence (visited on 18/11/2025).

[80] *Justice Department Leads Efforts Among Federal, International, and Private Sector Partners To Disrupt Covert Russian Government-Operated Social Media Bot Farm*. U.S. Department of Justice. 27th Aug. 2024. URL: https://www.justice.gov/usao-ndil/pr/justice-department-leads-efforts-among-federal-international-and-private-sector (visited on 18/11/2025).

[81] Daryna Khomycheva. *Russia Bets on Sovereign AI to Guard Culture from Western Digital Influence*. UNITED24 Media. 21st Oct. 2024. URL: https://united24media.com/latest-news/russia-bets-on-sovereign-ai-to-guard-culture-from-western-digital-influence-12765 (visited on 18/11/2025).

[82] Vadim Kozyulin. "Militarization of AI from a Russian Perspective". In: July 2019.

[83] Lancieri, Filippo, Edelson, Laura and Bechtold, Stefan. "AI Regulation: Competition, Arbitrage and Regulatory Capture". en. In: (2024). DOI: 10.3929/ETHZ-B-000708626. URL: http://hdl.handle.net/20.500.11850/708626.

[84] Christie Lawrence, Isaac Cui and Daniel Ho. "The Bureaucratic Challenge to AI Governance: An Empirical Assessment of Implementation at U.S. Federal Agencies". In: *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*. AIES '23. ACM, Aug. 2023, pp. 606–652. DOI: 10.1145/3600211.3604701. URL: http://dx.doi.org/10.1145/3600211.3604701.

[85] Maroussia Lévesque. "Smoke and Mirrors? AI Regulation and Corporate Power". In: *SSRN Electronic Journal* (2024). ISSN: 1556-5068. DOI: 10.2139/ssrn.4736131. URL: http://dx.doi.org/10.2139/ssrn.4736131.

[86] Zongjie Li et al. *An Empirical Study on Large Language Models in Accuracy and Robustness under Chinese Industrial Scenarios*. 2024. eprint: arXiv:2402.01723.

[87] M. Lott, C. Lally and L. K. Møller. *Stolen voices: Russia-aligned operation manipulates audio and images to impersonate experts*. Institute for Strategic Dialogue (ISD). 9th July 2024. URL: https://www.isdglobal.org/digital_dispatches/stolen-voices-russia-aligned-operation-manipulates-audio-and-images-to-impersonate-experts/ (visited on 18/11/2025).

[88] Office of Management and Budget. *Guidance for Regulation of Artificial Intelligence Applications (Memorandum M-21-06)*. Washington, D.C.: Executive Office of the President, 17th Nov. 2020. URL: https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf.

[89] Office of Management and Budget. *Memorandum for the Heads of Executive Departments and Agencies: Guidance for Regulation of Artificial Intelligence Applications (M-21-06)*. https://www.whitehouse.gov/wp-content/uploads/2020/11/M-21-06.pdf. Office of the President of the United States. Nov. 2020.

[90] Krystyna Marcinek and Eugeniu Han. *Russia's Asymmetric Response to 21st Century Strategic Competition: Robotization of the Armed Forces*. OCLC: 1374795107. Santa Monica, CA: RAND, 2023. ISBN: 978-1-9774-1067-2.

[91] Anna Nadibaidze. "Russia's Drive for AI: Do Deeds Match the Words?" In: *The Washington Quarterly* 47.4 (2024), pp. 137–154. DOI: 10.1080/0163660X.2024.2435162.

[92] Anna Nadibaidze. *Russian Perceptions of Military AI, Automation, and Autonomy*. English. Foreign Policy Research Institute, Jan. 2022.

[93] Giedrius Naprys and Jonas Ernestas. *"Tom Cruise" undermining Paris Olympics in Russian deepfake*. CyberNews. 29th Mar. 2024. URL: https://cybernews.com/news/fake-tom-cruise-undermining-paris-olympics/ (visited on 18/11/2025).

[94] Oleksii Nasiedkin et al. "Decoding manipulative narratives in cognitive warfare: a case study of the Russia-Ukraine conflict". In: *Frontiers in Artificial Intelligence* 7 (2024), p. 1566022. URL: https://www.frontiersin.org/articles/10.3389/frai.2025.1566022/full (visited on 18/11/2025).

[95] Joanna Okun-Kozlowicki. *Standards and Regulations: Measuring the Link to Goods Trade*. Tech. rep. Accessed: 2025-10-06. Office of Standards and Intellectual Property, U.S. Department of Commerce, 2016. URL: https://legacy.trade.gov/td/osip/documents/osip_standards_trade_full_paper.pdf.

[96] Littler Mendelson P.C. *2024 AI C-Suite Survey Report: Balancing Risk and Opportunity in AI Decision-Making*. Accessed: 2025-11-19. Littler Mendelson P.C., Sept. 2024. URL: https://www.littler.com/sites/default/files/2024_littler_ai_csuite_survey_report.pdf.

[97] Pat Pataranutaporn et al. *Synthetic Human Memories: AI-Edited Images and Videos Can Implant False Memories and Distort Recollection*. 2024. eprint: arXiv:2409.08895.

[98] *Personal Information Protection Law of the People's Republic of China*. Accessed: 2025-11-18. 2021. URL: https://en.wikipedia.org/wiki/Personal_Information_Protection_Law_of_the_People%27s_Republic_of_China.

[99] Maria Polovniuk. *Russian propaganda network Pravda tricks 33% of AI responses in 49 countries*. Euromaidan Press. 27th Mar. 2024. URL: https://euromaidanpress.com/2025/03/27/russian-propaganda-network-pravda-tricks-33-of-ai-responses-in-49-countries/ (visited on 18/11/2025).

[100] *Provisional Administrative Measures for Generative Artificial Intelligence Services*. Accessed: 2025-11-18. 2023. URL: https://www.sec.gov/Archives/edgar/data/1110646/000110465925034870/tm2512294d1_ex99-1.pdf.

[101] B. Renz and P. D'Anieri. *The AI Wave in Defence Innovation. O'Reilly*. en. 2024. URL: https://www.oreilly.com/library/view/the-ai-wave/9781000875010/xhtml/C_016_c8.xhtml.

[102] *Researchers Warn of Russian Disinformation Infecting AI Chatbots*. DISA. 20th Mar. 2024. URL: https://disa.org/researchers-warn-of-russian-disinformation-infecting-ai-chatbots/ (visited on 18/11/2025).

[103] Huw Roberts et al. "The Chinese approach to artificial intelligence: an analysis of policy, ethics, and regulation". In: *AI and SOCIETY* 36.1 (June 2020), pp. 59–77. ISSN: 1435-5655. DOI: 10.1007/s00146-020-00992-2. URL: http://dx.doi.org/10.1007/s00146-020-00992-2.

[104] Alexander Romanishyn, Olena Malytska and Vitaliy Goncharuk. "AI-driven disinformation: policy recommendations for democratic resilience". In: *Frontiers in Artificial Intelligence* 8 (July 2025). ISSN: 2624-8212. DOI: 10.3389/frai.2025.1569115. URL: http://dx.doi.org/10.3389/frai.2025.1569115.

[105] White House Office of Science and Technology Policy. *Blueprint for an AI Bill of Rights: Making Automated Systems Work for the American People*. 4th Oct. 2022. URL: https://bidenwhitehouse.archives.gov/ostp/ai-bill-of-rights/.

[106] Iskandar Sherqulov. "AI-Induced False Memories: New Research Shows 87 percent Success Rate in Memory Manipulation". In: (2025). DOI: 10.2139/ssrn.5142397. URL: http://dx.doi.org/10.2139/ssrn.5142397.

[107] C. Siegmann and M. Anderljung. *The brussels effect and artificial intelligence: How EU regulation will impact the global AI market*. en. arXiv. 2022. DOI: 10.48550/arXiv.2208.12645. URL: https://doi.org/10.48550/arXiv.2208.12645.

[108] Mona Sloane and Elena Wüllhorst. "A systematic review of regulatory strategies and transparency mandates in AI regulation in Europe, the United States, and Canada". In: *Data and; Policy* 7 (2025). ISSN: 2632-3249. DOI: 10.1017/dap.2024.54. URL: http://dx.doi.org/10.1017/dap.2024.54.

[109] Miodrag Soric. *Fact check: The deepfakes in the disinformation war between Russia and Ukraine*. DW. 18th Mar. 2022. URL: https://www.dw.com/en/fact-check-the-deepfakes-in-the-disinformation-war-between-russia-and-ukraine/a-61166433 (visited on 18/11/2025).

[110] Stanford HAI. *The 2025 AI Index Report*. https://hai.stanford.edu/ai-index/2025-ai-index-report. Accessed: 2025-09-20. 2025.

[111] State Council of the People's Republic of China. *AI Plus Action Plan*. State Council of the People's Republic of China. Beijing, Placeholder Month 2025. URL: Placeholder%20Official%20Chinese%20URL (visited on 18/11/2025).

[112] State Council of the People's Republic of China. *Full Translation: China's 'New Generation Artificial Intelligence Development Plan' (2017)*. Trans. by Rogier Creemers et al. Original document issued by China's State Council on July 20, 2017. Translation was updated in October 2018. New America Cybersecurity Initiative and DigiChina. 1st Aug. 2017. URL: https://www.newamerica.org/cybersecurity-initiative/digichina/blog/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/ (visited on 18/11/2025).

[113] State Council of the People's Republic of China. *New Generation Artificial Intelligence Development Plan*. Issued by the State Council (Guo Fa [2017] No. 35). State Council of the People's Republic of China. Beijing, 20th July 2017. URL: http://www.gov.cn/zhengce/content/2017-07/20/content_5211996.htm (visited on 18/11/2025).

[114] George J. Stigler. "The Theory of Economic Regulation". In: *The Bell Journal of Economics and Management Science* 2.1 (1971), pp. 3–21.

[115] Yuewei Sun. *The Impact of AI in the Chinese Workplace*. Research Paper. Chatham House, 18th July 2024. URL: https://www.chathamhouse.org/sites/default/files/2024-07/2024-07-18-workplace-ai-china-sun.pdf (visited on 18/11/2025).

[116] Elham Tabassi et al. *Artificial Intelligence Risk Management Framework (AI RMF 1.0)*. National Institute of Standards and Technology, 2023. DOI: 10.6028/NIST.AI.100-1. URL: https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf.

[117] *The Bletchley Declaration by Countries Attending the AI Safety Summit, 1-2 November 2023*. 2nd Nov. 2023. URL: https://www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration/the-bletchley-declaration-by-countries-attending-the-ai-safety-summit-1-2-november-2023 (visited on 20/09/2025).

[118] The Central People's Government of the People's Republic of China. *Placeholder Title of the Policy Document*. Please replace 'Placeholder Title of the Policy Document' with the actual title from the webpage. State Council of the People's Republic of China. 27th Aug. 2025. URL: https://english.www.gov.cn/policies/latestreleases/202508/27/content_WS68ae7976c6d0868f4e8f51a0.html (visited on 18/11/2025).

[119] The White House. *FACT SHEET: President Biden Issues Executive Order on Safe, Secure, and Trustworthy Artificial Intelligence*. https://bidenwhitehouse.archives.gov/briefing-room/statements-releases/2023/10/30/fact-sheet-president-biden-issues-executive-order-on-safe-secure-and-trustworthy-artificial-intelligence/. Accessed: 2025-09-20. Oct. 2023.

[120] Van Hong Tran et al. "Measuring Compliance with the California Consumer Privacy Act Over Space and Time". In: *Proceedings of the CHI Conference on Human Factors in Computing Systems*. CHI '24. ACM, May 2024, pp. 1–19. DOI: 10.1145/3613904.3642597. URL: http://dx.doi.org/10.1145/3613904.3642597.

[121] UNESCO. *Ethics of Artificial Intelligence*. https://www.unesco.org/en/artificial-intelligence/recommendation-ethics. Accessed: 2025-09-20.

[122] UNESCO. *Recommendation on the Ethics of Artificial Intelligence*. English. Programme and Meeting Document SHS/BIO/PI/2021/1. Adopted on 23 November 2021. Licensed under CC BY-NC-SA 3.0 IGO. Paris, France, 2022, p. 43. URL: https://unesdoc.unesco.org/ark:/48223/pf0000381137.

[123] James Vincent. *Putin says the nation that leads in AI 'will be the ruler of the world'*. The Verge. 4th Sept. 2017. URL: https://www.theverge.com/2017/9/4/16251226/russia-ai-putin-rule-the-world (visited on 28/09/2025).

[124] Morgan Wack et al. "Generative propaganda: Evidence of AI's impact from a state-backed disinformation campaign". In: *PNAS Nexus* 4.4 (Mar. 2025). Ed. by David Rand. ISSN: 2752-6542. DOI: 10.1093/pnasnexus/pgaf083. URL: http://dx.doi.org/10.1093/pnasnexus/pgaf083.

[125] Claudia Wallner, Simon Copeland and Antonio Giustozzi. *Russia, AI and the Future of Disinformation Warfare*. Emerging Insights. Royal United Services Institute (RUSI), 2025. URL: https://static.rusi.org/russia-ai-and-the-future-of-disinformation-warfare.pdf (visited on 19/11/2025).

[126] Wayne Wei Wang et al. "Artificial Intelligence "Law(s)" in China: Retrospect and Prospect". In: (2024). DOI: 10.2139/ssrn.5039316. URL: http://dx.doi.org/10.2139/ssrn.5039316.

[127] Graham Webster et al. *Full Translation: China's 'New Generation Artificial Intelligence Development Plan' (2017)*. Aug. 2017. URL: https://digichina.stanford.edu/work/full-translation-chinas-new-generation-artificial-intelligence-development-plan-2017/ (visited on 29/09/2025).

[128] Kevin Wei et al. "How Do AI Companies "Fine-Tune" Policy? Examining Regulatory Capture in AI Governance". In: *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society* 7 (Oct. 2024), pp. 1539–1555. ISSN: 3065-8365. DOI: 10.1609/aies.v7i1.31745. URL: http://dx.doi.org/10.1609/aies.v7i1.31745.

[129] Kevin Wei et al. *Managing Industry Influence in U.,S. AI Policy*. Research Brief RB-A3679-1. Available at https://www.rand.org/t/RBA3679-1. RAND Corporation, 2024. URL: https://www.rand.org/pubs/research_reports/RB-A3679-1.html.

[130] *Western AI Chatbots Susceptible to Russian Propaganda Influence: Study Findings*. DISA. 27th Mar. 2024. URL: https://disa.org/western-ai-chatbots-susceptible-to-russian-propaganda-influence-study-findings/ (visited on 18/11/2025).

[131] *William M. (Mac) Thornberry National Defense Authorization Act for Fiscal Year 2021. Pub. L. No. 116-283, 134 Stat. 3388*. https://www.congress.gov/116/plaws/publ283/PLAW-116publ283.pdf. 1st Jan. 2021.

[132] Claudia Wilson. *THE EU AI Act and Brussels Effect. How will American AI firms respond to General Purpose AI requirements?* Working Paper. Center for AI Policy, 2024, p. 8. URL: https://assets.caip.org/caip/EU%20AI%20Act%20-%20Brussels%20Effect.pdf (visited on 18/11/2025).

[133] Wanqiang Wu and Xifen Lin. "Access to technology, access to justice: China's artificial intelligence application in criminal proceedings". In: *International Journal of Law, Crime and Justice* 81 (2025), p. 100741. ISSN: 1756-0616. DOI: https://doi.org/10.1016/j.ijlcj.2025.100741. URL: https://www.sciencedirect.com/science/article/pii/S1756061625000175.

[134] Li Xiaoyan and Reynaldo Gacho Segumpan. "Navigating AI Adoption Challenges in China's Public Sector: Implications for Efficiency and Performance". In: *International Journal of Academic Research in Business and Social Sciences* 15.3 (Mar. 2025). ISSN: 2222-6990. DOI: 10.6007/ijarbss/v15-i3/25098. URL: http://dx.doi.org/10.6007/IJARBSS/v15-i3/25098.

[135] Xinhua. *China to deepen AI development, boost high-quality development: official*. 31st Jan. 2024. URL: https://english.news.cn/20240131/3d0169a3724446d18c6d96c.html (visited on 18/11/2025).

[136] Rui-Jie Yew and Brian Judge. "Anti-Regulatory AI: How "AI Safety" is Leveraged Against Regulatory Oversight". In: *Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. EAAMO '25. ACM, Nov. 2025, pp. 16–27. DOI: 10.1145/3757887.3763017. URL: http://dx.doi.org/10.1145/3757887.3763017.

[137] Brandy Zadrozny. "Russian disinformation sites linked to former Florida deputy sheriff, research finds". In: *NBC News* (29th May 2024). URL: https://www.nbcnews.com/news/us-news/fake-news-sites-florida-deputy-sheriff-russia-rcna154315 (visited on 19/11/2025).

[138] *Zakharova pointed out why Western countries want to monopolize AI technologies*. EADaily. 12th Aug. 2024. URL: https://eadaily.com/en/news/2025/08/12/zakharova-pointed-out-why-western-countries-want-to-monopolize-ai-technologies (visited on 18/11/2025).

[139] Angela Huyue Zhang. "The Promise and Perils of China's Regulation of Artificial Intelligence". In: *SSRN Electronic Journal* (2024). ISSN: 1556-5068. DOI: 10.2139/ssrn.4708676. URL: http://dx.doi.org/10.2139/ssrn.4708676.

[140] Jan Zilinsky et al. "Justifying an Invasion: When Is Disinformation Successful?" In: *Political Communication* 41.6 (May 2024), pp. 965–986. ISSN: 1091-7675. DOI: 10.1080/10584609.2024.2352483. URL: http://dx.doi.org/10.1080/10584609.2024.2352483.

[141] Mimi Zou and Lu Zhang. "Navigating China's regulatory approach to generative artificial intelligence and large language models". In: *Cambridge Forum on AI: Law and Governance* 1 (2025). ISSN: 3033-3733. DOI: 10.1017/cfl.2024.4. URL: http://dx.doi.org/10.1017/cfl.2024.4.

[142] Katarzyna Zysk. "High Hopes Amid Hard Realities: Defense AI in Russia". In: *The Very Long Game: 25 Case Studies on the Global State of Defense AI*. Ed. by Heiko Borchert, Torben Schütz and Joseph Verbovszky. Cham: Springer Nature Switzerland, 2024, pp. 353–374. ISBN: 978-3-031-58649-1. DOI: 10.1007/978-3-031-58649-1_16. URL: https://doi.org/10.1007/978-3-031-58649-1_16.

[143] Р. И. А. Новости. *Владимир Путин назвал Россию отдельной цивилизацией*. РИА Новости. Section: Новости. URL: https://ria.ru/20200517/1571580444.html (visited on 28/09/2025).

[144] 苟圣杰. "Research on the Construction of Government Regulatory System for Generative Artificial Intelligence". In: *Open Journal of Legal Science* 12.06 (2024), pp. 3670–3677. ISSN: 2329-7379. DOI: 10.12677/ojls.2024.126522. URL: http://dx.doi.org/10.12677/ojls.2024.126522.

Part II

The European AI Act

Chapter 7

Genesis

The genesis of the European Union's *Artificial Intelligence Act* was **a multi-year process** that evolved ethical and policy discussions into a legislation entirely dedicated to AI governance and its integration with European laws.

7.1 Early Foundations and Policy Discussions (2017–2020)

The concept of a mandatory legal framework for AI in Europe predates the official 2021 proposal, with mentions of the topic emerging as early as 2017, in the EU Parliament's resolution *Civil Law Rules on Robotics*[38](which, admittedly, did not bear long term fruits due to the criticism to the idea of giving a legal status, and liability, to autonomous robots[53]). Since then, the EU has been developing an integrated approach comprising policies, guidelines, and plans to tighten control over AI, particularly in terms of its safety and trustworthiness implications. Key efforts in this foundational period include:

- The **General Data Protection Regulation** (GDPR) of 2016 served as a significant precedent for the AI Act, profoundly influencing its foundational philosophy and structure with its **risk-based**, **human rights-focused** approach. The GDPR famously pioneered this model by requiring organisations to conduct a Data Protection Impact Assessment (**DPIA**), a direct precursor to the AI Act's Fundamental Rights Impact Assessment (**FRIA**). Research[106] shows a significant overlap between the conditions requiring a DPIA under GDPR and the high-risk categories in the AI Act's Annex III, which the AIA itself acknowledges by specifying that a FRIA is a complement to a DPIA if one has already been conducted for the same processing activities.

- The 2018 European Commission strategy, «*Artificial Intelligence for Europe*». This paved the way for the creation, also in 2018, of the *High-Level Expert Group on Artificial Intelligence* (HLEG), tasked to draft ethics guidelines. HLEG's work reached a milestone in April 2019, when it published the «*Ethics Guidelines for Trustworthy AI*» and the «*Policy and Investment Recommendations for Trustworthy AI*».

In 2020, the focus shifted from high-level principles to concrete planning:

- In February 2020, the **European Commission** published a **White Paper**[36] on AI. This document outlines a comprehensive plan to position Europe as a **global leader in research and innovation** in human-centric AI, standing on a regulatory framework that champions **trustworthy**, **safe** (for both fundamental and consumer rights) and **reliable** products. It also applies a **risk-based** approach while avoiding excessive burdens and uncertainty for innovators.

- In October 2020, a *parliamentary resolution*[44] and two *parliamentary reports*[42][43] pushed for the address of the topics of **ethical principles, civil liability, and intellectual property rights** in Artificial Intelligence, further preparing the ground for a dedicated legislation. Annexed recommendations [43] highlighted the need for citizens to feel **confident** that new technology will not cause harm to them, and the possible unpreparedness of existing laws to consider also immaterial harm done by AI and robotics.

By the end of 2020, the initial discourse on ethics and fundamental rights had evolved into a clear vision for a new regulation.

7.2 The Legislative Proposal and Negotiation Process (2021–2023)

In **April 2021**, the European Commission put forward its *Legislative Proposal* for the AI Act[40] («*Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts*»), marking the official start of an ordinary legislative procedure between the European Parliament and the Council of the EU to decide on the final text.

During the negotiation, a key topic of debate, raised by the need to define the **scope** of the regulation, was determining what constitutes an **AI system**[41]: both the Council and Parliament felt the need to change the initial definition, and, ultimately, they opted to align with the one given by the *Organisation for Economic Co-operation and Development*:

> An AI system is a machine-based system 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. Different AI systems vary in their levels of autonomy and adaptiveness after deployment[110].

The advent of **ChatGPT** and its competitors, in particular, required **integrations** almost overnight: the topic of closed-source foundation models emerged as a critical addition to be made and needed a new set of conceptual tools to regulate them, especially since it was reached the consensus that an AI system's objectives can be *implicit*[110], meaning, in case of Large Language Models, that they are not stated in the code, but acquired via imitation and reinforcement during training and use.

Deciding what **types** of AI should be banned entirely, and which were **exempted** from the scope of the act, was also a significant area of discussion [41]: Parliament heavily revised the initial list, and the Council made the critical addition of extending to

private actors the prohibition on using AI for social scoring; the Council added an explicit exclusion for security, defence and military applications and heavily modified the measures in support of innovation (especially regarding **Regulatory sandboxes**) and Parliament further added significant exemptions for **research** activities and **open-source** projects.

7.3 Final Agreement and Adoption (2023–2024)

The final form of the law was reached on 9 December 2023. The European Parliament formally adopted the EU Artificial Intelligence Act on **13 March 2024**, with a large majority of votes (523) in favour and 46 against; the opposition was led overwhelmingly by The Left group, with members from France, Spain, Portugal, Germany, Ireland, Belgium, and Czechia voting against the act, followed by the European Conservatives and Reformists Group (ECR), primarily MEPs from the Netherlands, Sweden, Croatia, Romania, and Slovakia. The remaining parties (e.g. PPE, Verts/ALE, Renew) were almost unanimously in favour of the act, with only a handful of dissenting votes[1][39].

[1] Among the Italian MEPs, there was overwhelming support for the Act across all political groups, with no voters against it and only seven abstained.

Chapter 8

Architecture and theoretical underpinnings to its approach

8.1 The Pyramid of Risk

Since the European Commission's initial proposal in 2021, AI systems have been organised in a **four-tiered** (*unacceptable*, *high*, *limited*, *minimal*) **hierarchy of risk**, which still underlies the Act's basic structure today. The strong priority is the identification and regulation of systems deemed «*high-risk*». Risk **categories** are based on their «*intensity and scope*»[1] of the risk that AI systems can generate through their operations, and frame the **set of obligations** they must satisfy.

8.1.1 Risk classification

8.1.1.1 Unacceptable risk (Prohibited practices)

The AI Act adopts a **risk-based** approach, identifying and **banning** a limited number of AI practices that are deemed to create an unacceptable risk to fundamental rights and Union values. While acknowledging the many beneficial uses of AI, the Regulation recognises that the technology can be misused for «*manipulative, exploitative and social control practices*». Such practices are considered particularly harmful because they go against the respect for human dignity, freedom, equality, and democracy enshrined in the *Charter of Fundamental Rights*.

The Act's Article 5 establishes an exhaustive list of AI applications that are prohibited, while further recitals offer context, justification, and clarification for these prohibitions. Although some **exceptions** exist, especially for law enforcement, these are tightly controlled with strong safeguards and oversight.

Manipulative or Deceptive AI Systems *Article 5(1)(a)* prohibits the placing on the **market**, putting into service, or **use** of AI systems that deploy:

- **Subliminal** techniques beyond a person's consciousness.
- Purposefully **manipulative** or deceptive techniques.

[1] Recital (26) AI Act.

This rule applies when the goal or result is to seriously distort how someone behaves, by making it much harder for them to make an **informed choice**. The manipulation must cause, or be likely to cause, significant harm to that person or someone else.

Context from Recitals add the following things to consider for an appropriate evaluation:

- **Subversion of Autonomy**: *Recital 29* explains that such techniques are prohibited because they can «subvert and impair» a person's autonomy, decision-making, and free choices. The harm can be material or immaterial, including **physical**, **psychological**, or **financial** damage. The techniques can be subliminal (e.g., audio or video stimuli beyond human perception) or otherwise deceptive in ways that people are not consciously aware of or are unable to resist.

- **Intent and Causality**: An **intention** to cause significant harm is **not required**; the prohibition applies as long as harm results from the manipulative practice. However, distortion resulting from external factors outside the provider's or deployer's control is not covered.

- **Exceptions**: The prohibition does not affect lawful and legitimate practices, such as **medical treatments** (e.g., psychological therapy or physical rehabilitation) conducted under applicable law and medical standards, or common commercial practices like advertising that comply with existing law.

Exploitation of Vulnerabilities *Article 5(1)(b)* bans AI systems that **exploit the vulnerabilities** of a person or a specific group due to their **age**, **disability**, or **social** or **economic situation**. This ban applies if the system is meant to, or actually does, seriously distort someone's behaviour in a way that causes, or is likely to cause, significant harm to them or someone else.

Recital 29 further elaborates that vulnerable groups can include persons living in extreme poverty or members of ethnic or religious **minorities**. The harm caused by exploiting these vulnerabilities may accumulate over time.

This prohibition complements existing consumer protection law, such as the *Unfair Commercial Practices Directive*, which already bans practices causing economic harm to consumers.

Social Scoring *Article 5(1)(c)* bans AI systems that **rate** or **classify** people or groups based on their **social behaviour** or **personal traits over time**. This kind of social scoring is not allowed if it leads to either of the following results:

- Detrimental or unfavourable treatment in **social contexts unrelated** to where the data was originally collected.

- Detrimental or unfavourable treatment that is **unjustified or disproportionate** to the social behaviour.

Recital 31 clarifies that social scoring may lead to «discriminatory outcomes and the exclusion of certain groups,» violating the rights to **dignity** and **non-discrimination**. The prohibition is designed to prevent AI systems from generating scores that result in

unjustified or disproportionate negative treatment. It does not, however, affect **lawful evaluation practices** of natural persons carried out for a specific purpose in accordance with Union and national law.

Risk Assessments for Criminal Offences *Article 5(1)(d)* bans AI systems that make risk assessments of natural persons to assess or predict the **risk** of them committing a **criminal** offence, when based only on profiling or judging their personality traits. There is an exception for AI that helps a human assess someone's involvement in a crime, as long as the assessment is based on clear, proven facts directly related to a crime.

Recital 42 anchors this prohibition in the principle of the **presumption of innocence**, stating that natural persons should be judged on their actual behaviour and not without a **reasonable suspicion** of criminal activity.

Untargeted Scraping for Facial Recognition Databases *Article 5(1)(e)* bans AI systems that build or add to facial recognition databases by collecting facial images from the internet or CCTV footage **without a specific target**.

Recital 43 explains that this practice is banned because it «adds to the feeling of **mass surveillance** and can lead to gross violations of fundamental rights, including the right to privacy».

Emotion Recognition in the Workplace and Education *Article 5(1)(f)* bans AI systems that try to read people's **emotions** at **work** or in **schools**. The only exception is if the system is used for medical or safety reasons.

On this matter, *Recital 44* mentions the «serious concerns about the **scientific basis** of AI systems aiming to identify or infer emotions», noting their limited reliability and the potential for discriminatory and intrusive outcomes. Given the inherent **power imbalance** in work and education, such systems could lead to detrimental treatment and are therefore prohibited.

Biometric Categorisation based on Sensitive Data *Article 5(1)(g)* bans biometric systems that **sort people** to guess their **race**, **political** views, **union** membership, **religion**, **beliefs**, **sex life**, or sexual orientation. This does not include lawful labelling or filtering of biometric data in law enforcement.

Recital 30 confirms this prohibition while clarifying its scope: it specifies that it does not prevent, for example, the sorting of images according to hair or eye colour in law enforcement contexts, as this is considered lawful labelling rather than inferring sensitive personal attributes. The notions of «*biometric data*», «*biometric identification*», and «*biometric categorisation*» are to be interpreted in line with existing data protection regulations like the GDPR.

***Real-Time* Remote Biometric Identification (RBI) in Publicly Accessible Spaces** *Article 5(1)(h)* generally bans the use of **real-time remote biometric identification** systems in public spaces for law enforcement. These systems quickly capture, compare, and identify people visiting a place, which can lead to constant surveillance.

The prohibition is subject to a set of narrowly defined exceptions where such use is strictly necessary, for example, for the targeted search for specific victims of serious crimes (abduction, trafficking, sexual exploitation) and missing persons, or the localisation or identification of a person suspected of committing a serious criminal offence. According to *Recital 38*, any processing of biometric data for purposes not included in these exceptional cases remains subject to the GDPR, and cannot be fully justified solely by the AI Act.

Recital 32 explicitly recognises that RBI is «*particularly intrusive*», may affect the private life of a large part of the population, and can dissuade the exercise of fundamental rights like **freedom of assembly**. Technical inaccuracies can also lead to biased and **discriminatory** results. For this reason, *Recital 33* further stresses that any exceptional use must be «*strictly necessary to achieve a substantial public interest, the importance of which outweighs the risks*».

A cornerstone of the safeguards applied to this case is the **requirement for prior authorisation**: *Article 5(3)* requires that each use must be authorised beforehand by a **judicial authority** or an independent **administrative authority** whose decision is binding. *Recital 35* explains this is to ensure **responsible and proportionate use**.

In any case, the use of these tools must be **limited in temporal, geographic, and personal scope** to what is strictly necessary. Furthermore, no adverse legal decision may be based solely on the output of the RBI system.

8.1.1.2 High risk

AI systems are considered high-risk if they have the potential to **significantly impact individuals' health, safety, or fundamental rights**. To identify them, the Regulation sets out two main routes:

- The first pathway covers AI systems that are **already being subject to safety regulation under the EU due to their sensitive nature** (for example, machinery, toys, lifts, medical devices, marine equipment, vehicles, appliances burning gaseous fuels, etc.). An AI system is labelled high-risk if it serves as a safety component, or if it is itself a product that needs an independent conformity check under those Union's harmonisation laws (listed in *Annex I* of the Act), even if the whole product or the AI component alone are not high-risk under those set of rules.

- The second pathway identifies specific stand-alone AI systems as high-risk based on their **intended purpose** in specifically defined, sensitive areas. AI systems used in these areas and their corresponding use-cases pose a significant potential for harm due to their impact on people's lives and rights. The areas, specified in *Annex III*, include:

  - **Biometrics**, including remote **biometric identification** systems, biometric categorisation based on **sensitive characteristics** (such as race, political opinions, or sexual orientation), and emotion recognition systems.
  - **Critical Infrastructure**: AI systems used as safety components in the management and operation of road **traffic** and the supply of **water**, **gas**, **heat-**

ing, and **electricity** are deemed high-risk due to the potential for large-scale harm if they fail.

- **Education and Vocational Training**: Systems used to determine **access** to educational institutions, **evaluate** learning outcomes, or **monitor** students during tests are classified as high-risk because they can significantly influence a person's educational and professional life course and may perpetuate discrimination.
- **Employment and Workers Management**: This includes AI used for **recruitment**, making decisions on **promotion** or **termination**, **allocating tasks**, and **monitoring** employee performance. These systems can have a profound impact on career prospects and workers' rights.
- **Access to Essential Services**: This category covers systems that evaluate **eligibility** for public assistance benefits, determine **credit** scores, are used for risk assessment in life and health **insurance**, and prioritise the dispatch of **emergency services**. The vulnerability of individuals dependent on these services justifies the high-risk classification.
- **Law Enforcement**: Systems used for assessing the **risk** of a person offending, evaluating the reliability of **evidence**, or **profiling** individuals in the course of criminal investigations are considered high-risk due to the inherent **power imbalance** and potential to infringe on fundamental procedural rights, such as the presumption of innocence.
- **Migration, Asylum, and Border Control**: AI used in this context, such as for risk assessments of individuals or examining **asylum and visa applications**, affects people in particularly vulnerable positions and is therefore classified as high-risk.
- **Administration of Justice and Democratic Processes**: This includes AI intended to assist judicial authorities in **legal interpretation** or to influence the outcome of an **election** or referendum, given the potential impact on the rule of law and individual freedoms.

Derogations The Regulation includes a specific derogation for systems that satisfy the characteristics as seen in *Annex III*, as they are considered not to pose a significant risk of harm to health, safety, or fundamental rights, because they cannot effectively influence the outcome of decision-making. This condition is met if the AI system is intended to:

- Perform a **narrow procedural task**.
- **Improve the result of a previously completed human activity**.
- **Detect decision-making patterns** without replacing or influencing human assessment.
- Perform «*a preparatory task to an assessment relevant for the purposes of the use cases listed in Annex III*».

Recital 53 clarifies these conditions with examples, such as an AI that formats data or improves the language in a previously drafted document (improving a human activity).

Derogations do not apply if the AI system performs **profiling of natural persons**; in such cases, the system is always considered high-risk. Providers who believe their system qualifies for derogation must document their assessment before placing it on the market.

8.1.1.3 «Limited» risk

An AI system is considered to have *limited risk* if it does not meet the criteria for any of the higher-risk categories. The provisions that interest these limited risk systems are found in *Chapter IV*, *Article 50*: «*Transparency obligations for providers and deployers of certain AI systems*», in which are cited, among others, «*AI Systems that Generate or Manipulate Content (Deepfakes and Synthetic Media)*».

The Act's philosophy is to allow these systems to be developed and used with minimal regulatory friction, promoting innovation and free movement within the internal market, while mitigating the «*specific risks of impersonation or deception*» that can arise, for example by imposing to distinguishably disclose the artificial origin of their outputs. For this reason, the AI Act imposes no new mandatory legal obligations, but *Recital 165* suggests providers **voluntarily** apply additional requirements related to environmental sustainability, AI literacy training, accessibility for persons with disabilities, stakeholder participation in the design and development process, and diversity within development teams.

8.1.2 The special case of GPAI models

While the pyramid is a valuable starting point, the final text of the AI Act had to be updated to adapt to the special nature of **GPAI models** , which are characterised by «*significant generality*», meaning they are not limited to a specific, narrowly designed purpose (but they can *generate* multiple specialised systems via fine-tuning). The change involves the introduction of a **parallel category**, with its additional set of obligations, that can span various levels of risk, and include an additional one: the *systemic* risk.

The result is a separate regulatory regime[2] organised in a two-tier approach:

- **Base obligations**[3] for all GPAI Model providers: **transparency** and **diligence**
    1. Drawing up and maintaining detailed technical **documentation** covering the model's training process, testing, and evaluation results;
    2. **Providing downstream companies** that **integrate** the model into their own AI systems with the **necessary information to understand its capabilities and limitations**, enabling them to comply with the Regulation. The minimum contents to be included are listed in *Annex XII*: «*a general description of the [...] model*» and «*description of the elements of the model and of the process for its development*».

[2] *Chapter V* of the AIA.
[3] Art. 53(1).

3. Establishing and upholding a policy to respect Union **copyright** law, including having a system put in place to identify and comply with reservations of rights by rightsholders.
4. Making publicly available a «*sufficiently detailed summary*» of the content used to **train** the model.

Open-source GPAIMs are **exempted** from obligations *(1)* and *(2)*, providing that they do not pose systemic risk.

- **Additional** requirements for GPAIMs posing a *systemic risk* (i.e. if the computational power used for its training exceeds a threshold of 10^{25} floating-point operations (FLOPs)[4], and/or if it meets the criteria of *Annex XIII* for the number of parameters of the model and its impact on the internal market)[5]:

  1. Performing, documenting and update comprehensive **model evaluations**, including adversarial testing (*red-teaming*), to identify and mitigate systemic risks.
  2. Continuously **assessing and mitigating** systemic risks that may arise from the model's development and use.
  3. Tracking, documenting, and reporting serious **incidents** to the AI Office without undue delay.
  4. Ensuring a high level of **cybersecurity** protection for the model and its physical infrastructure.

8.1.3 The reasoning behind a risk-based approach to new technologies[56, 107, 12]

The AI Act's tiered organisation was chosen for its balanced approach to regulating an ever-evolving technology that cannot be rigidly characterised in a future-proof manner, yet still needs to be addressed due to its significant potential to disrupt people's lives and rights. The so-called *precautionary approach*, which involves the outright banning of particular products from the market, remains applied only to cases that have been found to be without doubt excessively interfering with human rights.

The theoretical foundations of this approach are considerably more complex than the straightforward notion that laws should focus their attention on activities that pose the most significant harm. They come from questions about how risk is **conceptualised**, how regulatory systems can **adapt** to uncertainty, what institutional arrangements are needed to **sustain** risk governance, and how risk-based approaches **interact** with fundamental political and ethical issues. Understanding these deeper theoretical currents is essential for illuminating both the potential and the limitations of the approach.

At the most fundamental level, risk-based regulation emerges from a shift, already apparent in the 1960s, in how governments understand their role and duty towards their

[4] Art. 51(2).
[5] Art 55(1).

citizens;until the first half of the twentieth century, regulation was primarily reactive, aimed at making **restitution** for harm that had already been done. From then onwards, however, a new paradigm began to emerge, particularly in the United States and Europe, marking the birth of the *regulatory state*, in which the government was now responsible for **anticipating and preventing** future harms. This reorientation was partly philosophical[6] (reflecting a modern belief that risks can be expected and controlled) and partly practical, driven by the emergence of new technologies such as pharmaceuticals and pesticides, whose effects on health were clearly going to be long-latent and challenging to reverse. Within this larger context, *risk* became a key concept around which to construct and address regulatory problems in a certain way, rather than being the manifestation of the natural dangerousness of modern times.

8.1.3.1 Types of risk-based regulation

The literature on risk-based regulation distinguishes between at least two fundamentally different **paradigms** of how risk should be conceptualised and managed: the first is what scholars term the ***rational-instrumental*** model, which conceives risk as something that can be **objectively measured** through quantitative methods, assessed through formal risk assessment techniques, and **managed through technical intervention**. In this vision, risk is merely the product of probabilities and estimates that can be derived from historical data, laboratory tests, or models. The regulator's task, in this context, is that of the **technician** who should apply specialised knowledge to identify, classify systematically, and then intervene. The appeal of this approach indeed lies in the possibility for making decisions that appear more objective, defensible, and insulated from lobbying than political rhetoric or arbitrary judgement. However, this model has not avoided criticism from scholars who have found that it promotes a naïve understanding of the world, where empirical data can be sparse or altogether missing and some fundamental truths remain **irreducible into mathematics**. Indeed, a mathematical model requires a finite number of very selected parameters to be devisable and manageable, which implies beforehand **a selection of which risks to assess**, what thresholds to employ, and what events to count as significant. These choices all involve **politics** and are born of **confrontation** of different stakeholders, and cannot be resolved by technical analysis alone, as they require **judgements about values, priorities, and acceptable trade-offs that are inherently political**.

In response to these limitations, a second paradigm has emerged, which academics refer to as the ***deliberative-constitutive*** model of risk governance. This approach openly acknowledges that risk assessment necessarily involves wisdom and that decisions must be grounded not only in technical expertise but also in **dialogue** with affected stakeholders, consideration of social **values**, and explicit acknowledgement of the **uncertainties** and **preconceptions** embedded in risk assessments. This approach inserts risk *in* society, recognising that technical experts, policymakers, companies, and the gen-

[6] This can be inscribed in a bigger current that sees lawmaking as not just a description of natural and immutable facts of nature, but as the manifestation of a political *will* that prevails over many others (see part IV).

eral public form a plurality of perspectives; nowadays, what «normal» people believe to be risky is taken as input into the risk governance process together with expert assessment, meaning that effective risk governance now requires the ability to bring together technical analysis with democratic participation, to make explicit value choices and to ensure that those who will probably bear the consequences of the concretisation of a risk will have meaningful opportunity to **shape** those decisions.

The **tension between these two paradigms**, the *rational-instrumental* and the *deliberative-constitutive*, has profound effects on how the resulting regulation operates in practice and on the **perceived legitimacy of the regulator's intervention**: on the one hand, a purely technical regulator will be rejected on the grounds of its lack of consideration of the non-quantitative variables of the problem; on the other hand, relying too much on dialogue with stakeholders will expose his or her to regulatory capture and/or a wave of irrational discontent that will try to override sound technical judgement. The literature on the topic grapples with the delicate question of how to balance these two imperatives: how can regulators employ rigorous technical analysis while maintaining openness to political — and human — values? Existing work highlights that this balance is not easy to reach, and, once it has, it is not a *one-size-fits-all* solution valid forever: it needs **ongoing reflection and adaptation** as regulators learn from **experience** about what works in different contexts and social expectations evolve.

Implementing risk-based regulation is not merely a matter of preparing checklists and interview scripts, but also considering the **underlying implicit choices** made when drafting the law, even as the concrete risk landscape evolves. The scholarly literature makes clear that fundamental normative questions are at stake in every stage of the process: **which risks matter** and which can be ignored, **how to balance the goal of risk reduction** against other values such as innovation, economic growth or individual liberty, and **how to set acceptable risk thresholds** —that is, deciding how much residual risk is tolerable after action has been taken. The literature notes that these normative choices are often left **implicit** in regulatory discourse, although the AI Act has been among the virtuous exceptions, as some scholars argue that a prerequisite for legitimate risk-based regulation is excellent **transparency** and **robust democratic** processes. This remains true even *after* law deliberation: tasks such as assigning risk scores will always require **discretionary** judgements, even in highly quantitative systems, and require consideration of the firm's management, its commitment to regulatory compliance, and the reliability of data it provided.

Ensuring **consistency** in these judgements in the eyes of different inspectors and across time also requires robust **explainability**, documentation, calibration, and course-correction mechanisms. This reconnects to another concept, the one of the **politics of resource allocation**: as risk-based frameworks make explicit that it is not possible to protect against all risks, we must choose which risks to address intensively and which to leave under-resourced or unaddressed. This kind of prioritisation is much more defensible than arbitrary allocation, but it still creates political vulnerability for the one who chooses, because if a harmful event occurs, the regulator can be accused of negligence.

8.1.3.2 Effectiveness in different contexts

A second dimension of risk-based regulations concerns the question of whether they are appropriate for governing **all hazards and scenarios** or only some. In principle, risk assessment depends upon the availability of **reliable data** about adverse outcomes: if one cannot estimate probability with confidence, then he or she cannot either speak of *risk* in the technical sense, but instead of *uncertainty*. Academic literature on risk governance emphasises this distinction and presents different management approaches appropriate for each case, underlining how **applying risk-based regulation to *uncertain* problems may be misguided**.

The distinction described above is illustrated by the example of the stark contrast between well-studied hazards for which decades of empirical data have been accumulated (e.g. the safety of established pharmaceutical classes) *versus* those of novel technologies for which the long-term effects have not developed yet (such is the case for state-of-the-art AI systems): for well-understood risks, risk-based regulation has been functioning effectively as a systematic means of allocating resources according to observed patterns of harm.For highly uncertain domains that might not be the case, as jumping at talking about *risk* might lead regulators to **a false sense of security** as they believe that they have identified and measured the relevant hazards, when in fact **significant unknowns remain**. The literature suggests that in situations of **genuine deep uncertainty** alternative strategies may be more appropriate, such as the ***precautionary*** approach of restricting potentially harmful activities to **small, monitored environments** until they can be demonstrated to be safe, and **built-in** mechanisms for learning and course-correction as new details emerge. Both of these approaches are a fundamental part of the approach of the AI Act, a risk-based regulation at first approach, and a framework for handling uncertainty in the developing AI world at its core.

There are also **political and institutional preconditions** for risk-based regulation to function satisfactorily. The literature emphasises that risk-based regulation requires a «*political licence*» to operate, i.e. sufficient political support and **public trust**. In the past, during periods of political confidence in regulatory institutions, as existed in America and Europe in the 1990s and early 2000s, risk-based frameworks could be presented as **rational** responses to the need for efficient governance; however, the financial crisis of 2007-09 revealed the vulnerability of risk-based regulation when the underlying risk assessments proved to have been put into operation in a seriously flawed way. In that instance, regulators that had devised sophisticated risk-based frameworks nonetheless had failed to perceive **the systemic risks that fell outside existing models** and to **remain responsive** to changes in the broader institutional and political environment.

Finally, the literature increasingly emphasises the **synergy between risk-based regulation and other regulatory strategies** and the importance of understanding how risk-based approaches interact with alternative or complementary modes of governance. Risk-based regulation is often presented as a self-standing approach, but in practice, effective regulation typically requires the combination of multiple resources: clear **standards** that define minimum acceptable behaviour, a security-oriented company **culture**, and incentive-based instruments that **reward** compliance and **penalise**

non-compliance. The challenge is to understand how these different logics interact and enhance — or disrupt — each other. Truly effective risk governance requires regulators to be aware and ready to exploit these synergies, rather than assuming that a single risk-based technical tool can be applied out-of-the-box and in a standalone way.

In conclusion, risk-based regulation is neither a straightforward technical solution nor a fundamentally flawed approach, but rather a complex governance strategy whose success depends upon skilful navigation of multiple tensions and upon the willingness of regulators to engage in ongoing critical reflection.

The inconsistency with the risk-based approach of the general purpose models category In a recent publication in the *European Journal of Risk Regulation*, Prof. Martin Ebers [30] offers a critical analysis of the legal requirements for General Purpose AI (GPAI) models set by the AI Act. For him, GPAI models are not handled correctly using the risk-based approach of the AI Act, which also seems to give a misguided interpretation of the principle of technology neutrality. Instead of targeting actual risks arising from specific applications, the regulations focus on the underlying technology itself, and with prejudice typical of regulatory capture—a misalignment that, according to Embers, undermines both the effectiveness and fairness of the law.

In his work, Embers highlights a potentially critical gap between legislative obligations and practical reality, due to the challenge for original creators to predict all potential risks derived from the model's possible uses. Moreover, he considers the requirement for bias detection fundamentally flawed, arguing that bias is highly context-dependent and cannot be universally identified or resolved at the model level.

Another criticism is reserved for the unquantifiable obligations tied to the so-called *systemic* risks, as, by definition, systemic risks are not linked to specific use cases, making it arduous to identify or manage. Embers emphasises that the AI Act remains vague and fails to provide concrete guidance for providers on what counts as a systemic risk for GPAI or how it should be addressed, as it refers to much broader impacts on the market or society as a whole, making it a fundamentally new, and ambiguous, regulatory target.

The paper also challenges the AI Act's use of an arbitrary computational threshold: the presumption that any GPAI model trained with more than 10^{25} floating-point operations (FLOPs) **automatically** present a systemic risk. Embers argues that this standard is questionable for several reasons: the risk associated with a model also depends **strongly** on its underlying architecture and the quality of its training data, so much so that research shows that models trained using fewer FLOPs can be just as risky, even more than larger models. Prof. Ebers is so skeptical that he goes as far as to suspect that the FLOPs requirement has been set for purely *political* reasons: specifically, to exempt European start-ups like Mistral and Aleph Alpha from burdensome checks, while retaining the appearance of being grounded in empirical evidence.

8.1.3.3 Do first, regulate later?

The realm of AI is still in its nascent stage, and the numerous «twists and turns» that have been influencing the technology market in its infancy are still happening. Suddenly, one model or another comes out of the blue, dramatically altering the established rules overnight. In comparison with their inaugural iterations (for instance, ChatGPT was introduced at the conclusion of 2022), these services have undergone substantial modifications, a consequence of a regulatory intervention process that initially encountered difficulties in identifying appropriate definitions to contextualise these applications.

In a period of approximately three years, the domain of Large Language Models and foundation models has, to a considerable extent, stabilised around its leading proponents. These entities have assumed a dominant role in shaping market trends and the prevailing state of the art. However, Artificial Intelligence as it is currently conceptualised, especially in the domain of generative AI, is only now beginning to be the subject of extensive studies that are not exclusively technical and performance-oriented. To a certain extent, there is still a lack of knowledge regarding the effects it has on humans, especially when contact is deep, unfiltered and prolonged.

The "Regulatory Timing Problem" A. Robertson of the University of Chicago [108] shows how stalling to reduce mistakes and over-regulation, can **severely limit the regulators' options for intervention in the long run**: once enough time passes and a product becomes established, **an entrenched power structure** emerges that becomes «*extremely difficult to overcome, even if regulation—including regulating the product out of existence—is socially valuable*». Wansley [123] documents how this plays out in practice. Administrative agencies must satisfy high informational thresholds before acting; however, while regulators are acquiring information, the grassroots industry consolidates and becomes powerful, politically and economically. Then, «*[b]y the time agencies can justify regulation, the newly entrenched industries have the political capital to thwart them.*». Often, they are also already too **embedded in daily life** to consider blocking them without significant pushback from its users. In the end, delayed regulation has effectively enabled regulatory capture.

Such a phenomenon is not new and is not just a theoretical possibility: it has been deliberately exploited before with great success. Mazur and Serafin[74], in their article titled «*Stalling the State: How Digital Platforms Contribute to and Profit From Delays in the Enforcement and Adoption of Regulations*», show concrete examples of how digital platforms actively **slow the state down** by reinventing classifications, dragging out court proceedings, and generally being uncooperative, all in an effort to gain more time to earn as much money as possible, which translates in higher contracting power, and later protection from the State's intervention.

The AI Act's proactive and adaptive approach to regulatory design is once more a light in the dark. This kind of «*experimentalist regulation*»[123], in which independent authorities impose checks on risky emerging technologies while organising for protected environment to experiment with them, generates **fast, high quality evidence** for new legislation while **limiting exposure**, and keeps future regulatory options[123] open. This is for many a legitimate middle ground between foregoing present and future control and recklessy imposing arbitrary consequences, which have economic consequences that

persist even after a ban is lifted (e.g. Italy's ChatGPT ban caused[11]both an immediate market value loss (-6.8%) for related firms , and a lasting climate of uncertainty that discouraged economic development), and affect **predominantly smaller and younger firms**, further cementifying the supremacy of few, established superpowers.

Another side of this aspect is the AI Act's broad, flexible principle-setting, preferred to detailed rules: it allows regulators to adapt quickly as facts emerge, avoiding «*rules focused thickets*»[46]. A regulation that sets **high level objectives** (e.g. «*prevent discrimination*») allows for covering novel AI applications that were not considered when the law was drafted, avoiding the need for constant incremental updates as AI models evolve for new use cases. When needed, for the regulator it is only a matter of refining the principles as new information becomes available, without the long and formal rule-making process. Monitoring and, if necessary, sanctioning too become less focused on technicalities and more on actual outcomes (e.g., harm, bias) rather than on whether a specific clause was followed or checking line-by-line code compliance.

Of course,the problem for matter-of-fact, profit-chasing firms is to **interpret** the principles' broad language, which can create ambiguity and increase the cost of legal counsel or internal governance structures used to assess whether a product complies with abstract principles. Companies are also more limited in their regulatory capture effort[46], as broad principles are more complex for entrenched interest groups to game through loopholes, although they can still be shaped by lobbying regarding the wording of the principle itself, which can still dilute the its effectiveness.

As far as effective **oversight** is concerned, research has been exploring the fascinating world of *RegTech* [8] (short for *Regulatory Technology*), which refers to the **application of IT tools** (including *big data analytics*, *machine learning*, *sensor feeds* and *blockchain-based* solutions) to improve the **efficiency, accuracy and timeliness** of its processes. Through automated data collection, monitoring and reporting, RegTech is claimed to narrow the lag that separates the birth of a new activity and the start of effective oversight over it, giving regulators earlier insight into market developments and allowing them to act before entrenched dynamics solidify thanks to evidence-based, real-time alarms. Of course, as Bagby et al.[8] underline, while the promise of faster and more effective oversight is clear, challenges remain: it is not clear how this approach will mesh with the problem of **closed-source** AI models, which have been resisting since the beginning from external scrutiny .

In short, the full scenario depicted here fully supports the direction that the AI Act has taken from the outset: to embed a high-level principle and delegate detailed technical specifications to designated actors. While this requires periodic review cycles and revisions, the process is far from chaotic, as there are plenty of opportunities to inform it with data-driven insight from Reg-Tech tools and guide it with further interpretative guidance to mitigate compliance uncertainty, delivering both agility and accountability in the regulation of fast-moving technologies.

8.1.4 Risk assessment

While risk-based regulation is still considered the correct path forward for effective protection [30], researchers argue that the Act lacks crucial elements like a risk-benefit analysis and relies insufficiently on empirical evidence, particularly regarding the systemic risks of GPAI models: the AI act's parameters, namely the computational size rule and the general potentiality to affect public health, safety, security or fundamental rights, fail to provide a **quantitative** criteria for the probability of a harmful event and its severity, as is common in other risk-based product legislation (such as the *Medical Device Regulation*, which includes the provision for the explicit estimation of the probability and effective severity of the harm incurred, its detectability and the frequency of use of the product). Novelli et al. [79], for example, underline that the Act lacks a transparent, detailed **methodology for assessing risks in real-world situations**, relying instead on ambiguous **predictions** and **intents**. The Act also struggles to address risks to fundamental rights, such as privacy and fairness, because these concepts are difficult to define and apply consistently [92], leaving much room for **interpretation** in the hands of potentially uninformed actors who are not equipped to interpret the law fully.

Nevertheless, these flaws have been considered [30] to be fixable using the Act's existing tools, such as delegated acts and guidelines, in an effort to enforce a more genuine risk-based, proportionate implementation that relies more on real-world experience.

8.1.4.1 Proposed Methodologies for a « *Truly* » Risk-Based Approach

In academic and policy literature, several concrete proposals have been put forward to address these gaps and provide guidance on how to fulfil the risk assessment requirements of the Act. Key ideas include *scenario-based* methods [79], further *standard setting* [18], a *holistic documentation system* [52], and a stronger *liability* framework [64].

Scenario-Based Methodology Integrating the IPCC Framework Novelli et al. recommend a **scenario-based methodology** that focuses on assessing risk by examining specific, real-world situations rather than sorting AI systems by general application areas. Their approach integrates the **IPCC** (*Intergovernmental Panel on Climate Change*) **risk assessment framework**, which takes into consideration the interplay of risk **determinants** (such as exposure and vulnerability ì), the factors that create them, and the different kinds of risk that may arise (e.g.: physical, psychological, financial damage, loss of privacy, systemic bias, erosion of democratic processes, concentration of power, etc.). To refine their analysis, they advocate for a **proportionality** test to balance competing values and interests.

The benefits of this *semi-quantitative* framework include the opportunity to enable a more **nuanced categorisation**, allowing deployers of systems that are seemingly high-risk to challenge their classification by demonstrating low **actual** risk, and helping these deployers establish robust internal risk management systems.

A Standards-Based Approach: The CLAIRM Checklist While international standards, such as *ISO/IEC 23894* on AI risk management, already exist, the team at the *European Commission's Joint Research Centre* (JRC) considers them to be too

high-level and non-prescriptive to fully meet the AI Act's specific needs, particularly in terms of risks to fundamental rights.

To fill this gap, a new **European standardisation initiative** is underway, proposed by the *CEN-CENELEC Joint Technical Committee 21* (JTC 21). This initiative, the *Checklist for AI Risk Management* (*CLAIRM*)[18], is intended to be a practical and prescriptive European norm that will provide a **granular set of technical requirements** to guide the management of AI risks. The document aims to identify and describe specific sources of risk, outline the potential harms that could arise, and recommend **concrete countermeasures**, with the ultimate goal of helping AI providers to select and implement easily the most appropriate risk mitigation measures for their systems and establish a European-level reference.

As of today, the standard is still in its Pre-draft stage [[3]].

"*AI Cards*" for Risk Documentation and Management *AI Cards*[52] is a framework designed to make **documenting and sharing information about AI systems** easier and more transparent by organising details on how the system is meant to be used, its technical background, and how risks are managed in a clear and **structured** way.

What sets this approach apart from others that employ long, natural language descriptions and reports is that it uses two formats to share information: one, human-readable, provides a transparent and easily understandable overview of the AI's key aspects; the other, **machine-readable**, can be combined with standard web technologies made for interoperable querying (e.g. the SPARQL Protocol and RDF Query Language) [103] to be freely exchanged along the AI usage chain.The machine-readable format is significant for risk assessment procedures, as it facilitates the **update** of information as the AI system evolves, helps identifying what needs to change when new legal rules emerge, and enables the development of automatic tools for compliance checking.

Strengthening the Liability Framework Kretschmer et al.'s argument [64] is that the **ex-ante** risk-based approach is fundamentally ill-suited for a versatile and dynamic technology like AI. Instead, they suggest that a strong liability system that establishes crearly who is responsible after harm happens would **remove most of the complexity of the preventative approach** (which requires qualities that are statistically and practically impossible to reach,such as complete and error-free training datasets, and meaningful human oversight) and give clearer and better **incentives** for transparent systems. The goal is to motivate developers to design models that are transparent, auditable, and capable of rapid correction, and to incentivise deployers to maintain vigilant oversight of their applications, implement robust safeguards, and establish productive communication channels with developers in order to detect and report emerging risks.

The core of the proposal is a **liability matrix** that distinguishes between different causes of harm: *exogenous* (external causes that are outside the direct control of users or providers, such as hallucinations) and *endogenous* (i.e. resulting from the actions of providers or users, such as data poisoning or prompt injection). This approach has the advantage of better **aligning responsibility with the party best positioned to mitigate** a specific risk (see Table 8.1).

The proposal for joint liability in the case of continuously deployed systems facing endogenous risks represents a pragmatic solution that accepts the inherent unpredict-

Harm	Exogenous	Endogenous
Deployment		
One-Shot	The **developer** is responsible for updating the static model when the external world changes	The **deployer** is responsible for mitigating misuse
Continuous	The developer of the core model is responsible for managing ongoing external changes	**Both** the model **developer** and the product **deployer** share responsibility for harms from strategic manipulation, as both influence the system's design and its real-world interaction.

Table 8.1: Liability matrix according to Kretschmer et al.

ability of advanced AI systems and replaces an impossible demand for perfect *ex-ante* safety with a compelling **incentive for rapid, collaborative remediation** of unlawful outcomes. Without a clear rule for joint liability, blame risks to be shifted back and forth ; for example, the deployer might argue that harm resulted from flaws in the developer's model, claiming the exploit was unpredictable and the model too easy to manipulate. On the other hand, the developer may assert that their model is a general-purpose tool, and that the deployer should have implemented appropriate safeguards, filters, and monitoring for their specific application. Ultimately, **the person harmed** has no clear way to seek redress, as each side blames the other, leaving them with no recourse. With Joint Liability, instead, the harmed user can take action against both companies, forcing both to immediately take the system offline and issue a patch that neutralises the vulnerability, and stopping either from stalling in the hope of not being the one that will be forced to act to fix the problem.

Rather than simply dividing blame, the approach is intended to **foster cooperation and timely remediation** when issues arise, as the base understanding is that **neither party can manage all risks independently**: developers possess the expertise required to address vulnerabilities within the core model, while deployers have access to real-time operational data and context to identify harm as it occurs. By holding both parties jointly liable, the framework promises to encourage them to combine their respective strengths to address problems effectively and promptly, especially since it is unrealistic to expect any deployed system to be utterly immune to novel exploits from the get-go.

8.2 A human-centric approach

The European Commission's Communication (*COM(2019) 168 final*) «*Building Trust in Human-Centric Artificial Intelligence*»[35] advocates for *Human-Centric Artificial Intelligence* to foster societal trust and enhance Europe's competitive position in the AI market. The initiative seeks to mobilise at least 20 billion euros annually in AI investments over the next decade, with 1 billion euros per year provided by Horizon Europe and Digital Europe.

This document is a pivotal moment in the establishment of the human-centric approach as a wayfinder for the AI Act, noting that ensuring trustworthiness is essential, as it necessitates that AI systems adhere to rigorous, legal, and ethical standards.

The AI high-level expert group that handled its writing initially outlined **seven key requirements**, including *human agency*, *privacy*, *fairness*, and *accountability*, paving the way for the Commission to initiate a piloting phase to evaluate how these ethical guidelines function in real-world scenarios, with the aim of using lessons learned to refine the trustworthy AI framework.

8.2.1 The Seven Requirements

These guidelines outline seven interconnected requirements that form **the foundation for trustworthy and people-centred Artificial Intelligence**. Taken together, these seven requirements want to establish a comprehensive framework, inserted in the solid tradition of EU Regulations and Acts on adjacent themes (e.g. GDPR, Cybersecurity

Act), that integrates legal compliance, ethical principles, and technical robustness, for **trustworthy Artificial Intelligence**.

- The first pillar is **human agency and oversight**: Artificial Intelligence systems should function as **enablers**, **supporting** individuals in making **informed** decisions that align with their goals and values. To ensure this, mechanisms such as *human-in-command* are recommended, enabling users or public authorities to intervene on, **override**, or **suspend** automated decisions when required.
  As the traditional approach of «*Human-in-the-loop*», which involves a human operator's intervention at every decision step, becomes more and more impossible and even undesirable due to its natural slowness and error-making. In this proposal, the human operator can still intervene on, override and monitor all AI activities, while allowing for true automation.

- **Technical robustness and safety** represent the second requirement. Trustworthy Artificial Intelligence must remain reliable, secure, and resilient throughout its life cycle, effectively managing **errors**, inconsistencies, or malicious **attacks** without causing harm. This involves implementing ***safety-by-design***, establishing **fallback** procedures, and ensuring that outcomes are **reproducible** and accurately reflect the system's knowledge.
  Safety-by-design remains here a fundamental piece of the EU's approach to technology, requiring that safety considerations be built into a product or system from the **earliest design stages** rather than added as an after thought. By embedding fail-safe defaults and robust testing and ongoing monitoring, safety by design stresses the need to protect against motivated attackers and to ensure that systems remain safe even when components fail. This principle underpins the AI Act's whole mandatory risk-management system, which must continuously identify, assess, and mitigate harms to health, safety, and fundamental rights throughout the lifecycle.

- The third pillar, **privacy and data governance**, requires that personal data be protected at all stages of processing. For individuals to **retain control** over their data, Artificial Intelligence systems should prevent the disclosure of sensitive attributes about the people on whom it was trained. High-quality, bias-free datasets are essential, and the integrity of training data must be maintained through rigorous testing, thorough documentation, and controlled access. High quality data are essential for trustworthy outcomes, and their scarcity is a systemic problem with roots in multiple sources such as shady data gathering, data brokering, profilation cookies, and many more; these phenomena are so widespread outside the EU (and, in a milder form, also inside it) that the problem of the **sanitation of the data sources** cannot be ignored since the inception of any design stage, primarily since it represents the knowledge base of the future AI Model.

- **Transparency** constitutes the fourth requirement. Complete traceability of Artificial Intelligence decisions and its underlying development process, including data collection, labelling, algorithmic design, and deployment rationale, must be documented and made accessible to **independent auditing**. Explainability should reach every stakeholder in a way tailored to their understanding and role, and users

must be clearly informed of the risks and terms of their usage.
While not explicit in the *Communication*, there is a strong **tension between transparency and data privacy**.The balance between privacy, usefulness and transparency of training data requires careful delineation of what is supposed to remain private, what can become part of the model's knowledge, and in what ways (for example: information on the classification process could help an attacker to infer the original data just by seeing the labelling the model applied to them).
Moreover, the dataset used for training is often seen by developers of AI models as a **key asset** to be competitive, and as such, it is kept as a *de facto* secret to maintain market advantage. The same can be said for any kind of technical document that can reveal the inner workings of the latest, state-of-the-art model that was the fruit of huge monetary investments.

- Without transparency, there is no **accountability**, another requirement. Precise mechanisms must be established to assign **responsibility** for AI systems and their outcomes before and after deployment. **Auditable** processes, as well as internal and external reviews, are essential. Accessible redress pathways must be provided. These ensure that any adverse impacts can be identified, documented, and remedied fairly.

- **Diversity, non-discrimination, and fairness** comprise the sixth pillar. Building upon the third pillar, datasets and algorithmic designs must be assessed for historical biases, incompleteness, or discriminatory patterns that could result in unfair treatment. Engaging diverse design teams and consulting affected stakeholders throughout the system's life cycle are two suggestions on how Artificial Intelligence can better serve individuals of all types and backgrounds.
Inclusivity is addressed in the two ways bias can creep into an AI system: the diversity of data sources and the universal design of its algorithm.

- This last pillar is closely linked to the sixth requirement: **societal and environmental well-being**. Artificial Intelligence should be developed and deployed to support biodiversity and protect the environment for current and future generations, and the broader social impacts of technology on democratic processes and society must be carefully evaluated and mitigated as necessary.
In particular, it is imperative to extend beyond the measurement of operational emissions to include carbon emissions from hardware manufacturing and the resource consumption involved.

As usual, the degree with which these pillars are adopted, especially when the choice comes with *trade-offs* (e.g. privacy and transparency, human agency and automatic, lightning-fast responses to threats), is not always easy to decide. A test phase after publication was considered to assess the requirements' reception among a wide range of stakeholders, including public administrations and private firms, with a host of consultation activities to provide opportunities to give their contribution. Complementary programmes such as *AI Alliance*, *AI4EU*, and networks of AI research excellence centres have also joined the challenge of the development, testing and deployment of trustworthy AI.Collaboration is key up to the highest organisational levels: in the Communication it is explicitly stated the goal of the EU to shape international AI ethics by cooperating with

«like-minded» countries, contributing to multilateral fora (G7, G20) and standardisation bodies, in an effort to spread the *gospel* of «human-centric AI» globally, leveraging Europe's reputation for safe, high-quality products.

8.2.2 The ethical guidelines approach

In 2019 [57], the European Commission's *High-Level Expert Group on AI* (AI HLEG), established in June 2018, published the latest revised version of the *Ethics Guidelines for Trustworthy AI*. The central thesis asserts that **Trustworthy AI** necessitates three essential, though individually insufficient, components: **lawful, ethical, and robust operation**.

The framework directs stakeholders beyond abstract principles by emphasising the practical implementation of ethical and robust AI across these three layers. It presents foundational ethical principles, such as human autonomy, fairness, and the prevention of harm, specifies seven key requirements, and introduces a **Trustworthy AI assessment list**, but leaves out explicit legal compliance guidance.

The three pillars of Trustworthy AI, lawfulness, ethics, and robustness, were designed to function **collaboratively** throughout the entire AI system life cycle, rather than as isolated requirements. Their interaction is complex and can assume diverse shades depending on context, but, in general terms, it can be described as follows:

1. **Lawfulness as the foundation of all actions**: AI developers must first ensure that the system complies with all relevant EU primary and secondary legislation (such as the GDPR, product liability rules, and anti-discrimination directives) as well as any sector-specific regulations from the design phase onwards, before even writing a single line of AI-enabled code.

2. **Ethical guidance during all steps should extend beyond legal requirements**: after legal compliance is established, ethical considerations **further inform** system design to include requirements that are not translatable in clear-cut numerical parameters or schematic checklists: respecting human autonomy, preventing harm, ensuring fairness and explicability, and upholding fundamental rights, even in areas where legislation is absent or outdated, can only come with human, informed analysis of the non–quantitative world. As these decisions covertly influence data governance policies, user interaction design, and stakeholder participation processes, they remain critical for societal acceptance and benefit.

3. **Robustness is not only a technical requirement, but a foundation for lawfulness and ethics**: it includes data security, algorithmic and operational safety, reliability, resilience to attacks, and fallback mechanisms—ensuring that the system operates as intended and avoids unintended harm. In this, it supports both legal obligations such as safety standards and accountability for damages and ethical imperatives such as harm prevention and fairness through reliable outcomes .

Depending on the actual development phase of the system, the three pillars will appear differently. During the conception and design stage, the steps will be enacted, for example, by a clear indication of the legal bases for data processing, a fundamental rights

impact assessment, the application of **privacy-by-design** techniques and the creation of robust data pipelines that have been tested for leakages and rogue usage, all documented for further inspection also down the line. During development and testing, it will be imperative to document everything as evidence for ongoing compliance, including the **evaluation** of fairness, error and hallucination metrics, the result of adversarial **testing**. During deployment, auditing trails will be continuously kept updated as part of **monitoring** operations, to allow for **oversight** by humans and the reconstruction of the algorithm's decision path, eventually triggering fallback plans and incident response procedures that align with the legal obligations.

The approach of treating lawfulness, ethics, and robustness as intersecting layers rather than sequential steps enables practitioners to effectively and continuously **verify** that measures taken for one pillar reinforce the others, leading to AI systems that are compliant, value-aligned, and safe in real-world operation. Nevertheless, the guidelines themselves acknowledge in their content that the three components may also *conflict*. For example, a feature that maximises *predictive accuracy* (an ethical goal) could raise *privacy* concerns (lawfulness goal) or increase *vulnerability* to adversarial attacks (robustness goal). In such situations, conflicts should first be **identified** through a systematic impact assessment; then, the necessary balance and trade-off assessment for the selected path should be **documented**, with the involvement of relevant stakeholders. A good solution attempts to **minimise harm**, typically by incorporating human oversight, adjusting the algorithm, and/or providing user-friendly **opt-out** mechanisms. **Continuous assessment** is a crucial practice in implementing human-centric AI systems.

8.2.2.1 The importance of balance and context

The *Ethics Guidelines for Trustworthy AI* **do not, in this case, provide an explicit high-risk versus low-risk categorisation framework** with detailed tier-specific adjustments, which was done extensively in other settings (the AI Act). However, it still offered essential guidance on how **proportionality** and **context** should shape the assessment, which is not intended as a one-size-fits-all checklist, but rather a list of considerations that should be graded for relevance to the actual case. The assessment should be tailored to match the magnitude and type of risk posed by the system with respect to the surrounding environment, i.e., the severity of potential consequences in that context (consider, for example, the consequences of an error for a shopping recommendation engine versus an algorithm for medical diagnosis). For **high-risk** applications (directly affecting fundamental rights, safety-critical), *technical robustness and safety* means carefully assessing security vulnerabilities, the impact of failures, and accuracy testing, while *auditability* means the chance for independent, third-party inspection and comprehensive documentation on all phases of the lifecycle; finally, *transparency* will mean timely and open explanations , the candid disclosure of risks in the model's usage, and explicit communication of its limitations. On the contrary, for applications with much lower risk and minimal impact on individuals, *explicability* may rely on simpler documentation rather than real-time explanation interfaces, *oversight* may be periodic rather than continuous (e.g. just HOTL mechanisms to intervene as needed), and *testing* may follow standard software-quality metrics rather than adversarial red-teaming.

8.2.3 The ecological impact of AI

The environmental impact of Artificial Intelligence, particularly in relation to the vast, resource-hungry Large Language Models, has become a heartfelt, but still relatively ignored point of discussion when talking about responsible AI development, as this dramatic escalation in model complexity directly correlates with unprecedented computational demands, transforming AI from a research curiosity into a sector with measurable ecological consequences. This issue encompasses both the considerable resource costs associated with training and deploying AI systems, as well as the substantial benefits that AI applications can offer in addressing climate-related challenges. The growing environmental footprint of AI has consequently prompted increased transparency and more robust interventions.

Central to these concerns are the direct environmental costs, often referred to as the AI footprint, which stem from the scale and complexity of foundation models and result in significant energy consumption, resource depletion, and carbon emissions throughout their entire lifecycle. Emissions data further illustrate the variability of environmental impact across different models, with training emissions for leading AI systems that are substantial and documented: training GPT-3 generated approximately 500 metric tons of carbon dioxide equivalent, comparable to driving an average passenger car for about one million miles[45]; after training, inference operations for GPT-3 were estimated to produce approximately 12,800 metric tons of CO_2 annually, roughly 25 times the emissions of the training phase. To further contextualise this impact, the energy consumed to train and operate ChatGPT has been estimated to equal the annual carbon emissions of 175000 Danish citizens[24]. Another study [111] has shown that many AI models do not even report carbon emissions.

In May 2024, Nature declared that the direct impacts of AI on climate so far are relatively small, as it reported an estimation attributing to AI only 0.01% of all global greenhouse emissions [68]. These results appear mild, especially when considering other significant sources of pollution present nowadays. Nevertheless, it is essential to contextualise the data : first, the field of AI emissions is today less studied than other industries, with fewer standards for assessment and less know-how on the opportunities for optimisation; second, the fact that the AI boom is still in its initial phases means that current consumption might increase exponentially once the technology has had time to grow further, as forecasts by the International Energy Agency[58] indicate that the AI industry alone will increase its electricity demand ten fold by 2026 with respect to 2023 , with the demand for AI services expected to rise by 30–40% annually over the next 5–10 years.Moreover, beyond the obvious operational and training energy expenditure, the environmental implications of AI extend into the realm of the manufacture of essential components such as **GPUs**, which is associated with the extraction and processing of raw materials (**rare earths** (REEs) in particular, but also other, often hazardous for human health, substances including tungsten, palladium, cobalt, mercury, lead and tantalum[61]), an activity that contributes to ecosystem degradation (e.g. water contamination, soil erosion) and pollution: mining typically involves open pit extraction, extensive use of acids and large volumes of water, leading to the release of hazardous tailings into rivers and groundwater[[47], [73]]. The cost is even higher if we include the fact that the extraction of REEs is geographically concentrated in the Democratic Republic of Congo and China, often under conditions of child labour, unsafe working

environments and limited economic benefit for host communities[73]. Fuelling this market means encouraging global inequality in which the regions that supply the critical minerals receive **little of the economic upside from the AI boom, while bearing all its burdens**.

Even if extraction were to be carried out ethically and with minimal environmental disruption, semiconductor production processes remain among the most energy intensive industrial activities, in addition to emitting dangerous fluorinated gases. Reports[65] show how an Intel semiconductor fabrication facility in Arizona used 927 million gallons (3.509.000 cubic meters) of potable water and produced ca. 15 000 tonnes of waste in a single three month period. Then, after production, the operation of data centres, which are indispensable for running large-scale AI models, implies high water consumption due to the cooling requirements of high-performance computing, thereby aggravating water scarcity in vulnerable regions and disrupting temperature-sensitive ecosystems.

Ultimately, the continuous production also indirectly fuels the growing challenge of **electronic waste handling** in light of rapid technological obsolescence and degradation from intensive use. The rapid turnover of AI-optimized hardware intensifies the waste stream, as GPUs launched to meet the latest model sizes are often retired within a few years, creating garbage that is exported to the Global South[2], where informal recycling releases substances such as lead and mercury[54] to the environment and **to the human workers**. Forecasts[122] predict a potential accumulation of between 1.2 and5.0 million tonnes of e-waste between 2020 and 2030 if current disposal practices continue, which is a sign of future crisis as the sheer volume of discarded chips is already beginning to strain existing recycling infrastructure[15]. This infrastructure, as of today, still uses conventional methods that recover only a fraction of the valuable REEs, so it is imperative to start endorsing more circular economy approaches, and prioritise **repairability, upgradeability and recyclability** in product design[54].

Studies made to address these complex and interrelated impacts have shown the need for a multifaceted approach that encompasses **technical, operational, regulatory, and governance strategies**: from enhancing energy efficiency through **software optimisation** to the strategic selection of locations and **sustainable energy sources**, favouring regions with cleaner energy grids to minimise carbon emissions. As the *Green AI* research field matures considerably, some studies are demonstrating energy savings exceeding 50%, and in some cases reaching 115%, through **optimisation** of hyperparameters and **architectural** choices[119]. However, recent scientific literature indicates that this alone will prove insufficient to ensure generative AI sustainability, because other issues, mainly future AI development policies and rare metal utilisation, will alone raise the problems of our Planet's finite resources[10].

8.2.3.1 The AI Act's perspective

The EU AI Act incorporates environmental and societal well-being as key ethical principles for trustworthy AI, so it also encourages a shift towards greater energy efficiency.

The Artificial Intelligence Act explicitly links AI regulation to environmental sustainability in several places, starting from its very first Recital:

> The purpose of this Regulation is to [...] promote the uptake of human centric and trustworthy Artificial Intelligence (AI) while ensuring a high level of [...] environmental protection.

Similarly, Recital 27 subscribes to the Ethics Guidelines described in 8.2.2:

> [...] In those guidelines, the AI HLEG developed seven non-binding ethical principles for AI which are intended to help ensure that AI is trustworthy and ethically sound. [...] Without prejudice to the legally binding requirements of this Regulation and any other applicable Union law, those guidelines contribute to the design of coherent, trustworthy and human-centric AI, in line with the Charter and with the values on which the Union is founded. [...] Social and environmental well-being means that AI systems are developed and used in a sustainable and environmentally friendly manner [...].

In concrete terms, this translates to a number of provisions that focus on transparency, documentation, and periodic review, rather than imposing explicit, across-the-board numeric limits or mandatory carbon budgets for AI training/deployment. This was not the case in the earlier versions of the AI Act, as the so-called *trilogue version* removed [95] most of the detailed rules, with only the core clauses remaining in the final version :

- Energy consumption **logging**: *Annex XI* (the documentation requirements for general-purpose models) explicitly requires providers to include «known or estimated energy consumption of the model», while *Article 11* includes among the instructions for use «the computational and hardware resources needed», from which it is possible to infer energy consumption.
- **Fundamental-rights impact assessment**: Environmental protection is recognised as a fundamental right by Article 37 of the *Charter of Fundamental Rights of the European Union* (which the AI Act uses explicitly as one of its foundational texts), so, a Fundamental Rights Impact Assessment should also consider the foreseeable adverse impacts on the environment.
- A high level of protection of the environment, including «*protection of biodiversity, protection against pollution, green transition measures, climate change mitigation and adaptation measures*» (*Art. 59*) is kept as a requisite for even the most cutting-edge cases belonging to special ***sandboxes*** [7] (see 8.4 on regulatory sandboxes).
- **Standardisation** for **reporting**: The text retains an obligation (*Art 40*) for the Commission to commission standards on documentation about an AI-system's resource performance, covering the resources used during both training and deployment.

As of today, the greatest support in complying in practice so comes from a growing body of **technical norms that establish standardised measurement protocols for emissions**, (see, for example, the tech report ISO/IEC TR 20226:2025 (*Environmental sustainability aspects of AI systems*)).

---

[7]

– *Functionally separate, isolated and protected data processing environments* (Art 59(1(a(ii))))

8.3 The problem of General Purpose models

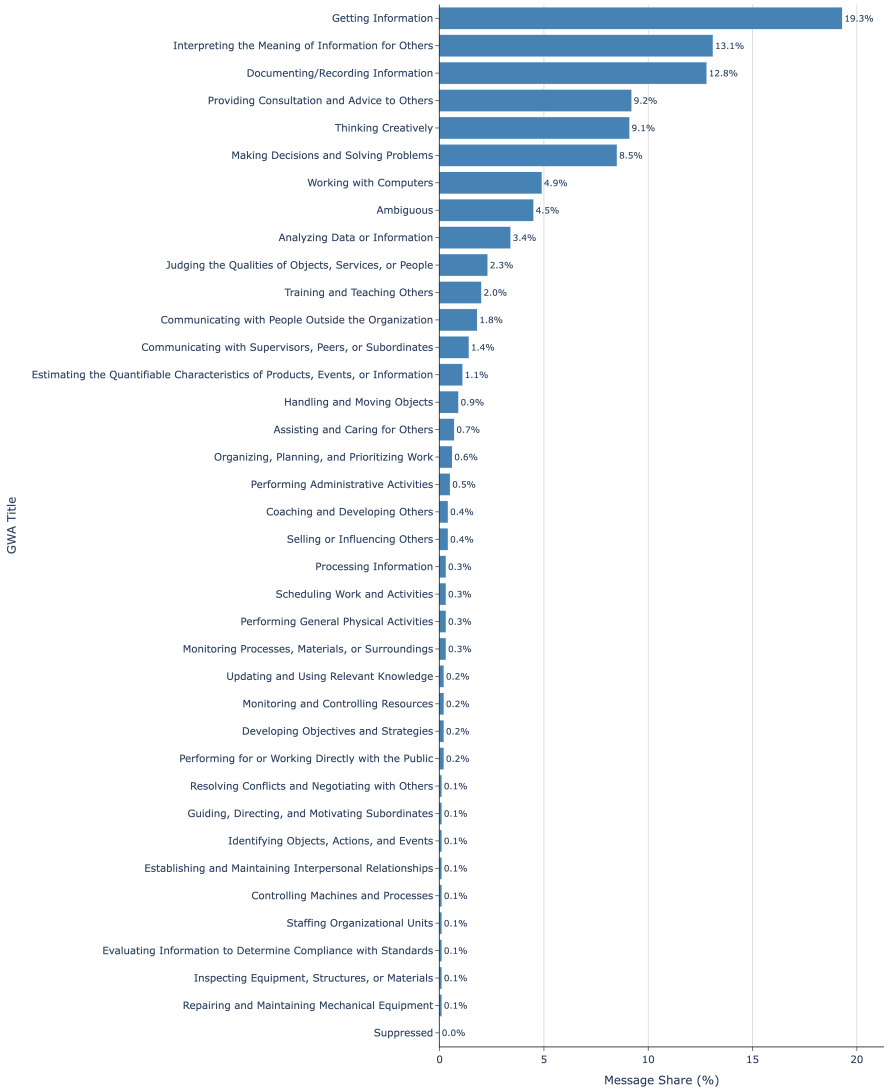
A *Foundation Model* is an AI model that is trained on broad, extensive data, often using self-supervision due to the massive scale of its operations [112]; since it was not built for one specific task, it serves as a base, or "foundation", for other AI models that are specialised through methods like *fine-tuning* or *prompting*. Prominent examples of foundation models include the GPT family and Google's BERT, and all Large Language Models (LLM).

The general-purpose nature of foundation models comes from their ability to perform *emergence*: this term refers to the way a model's capabilities are **implicitly induced from** the model's large-scale **training rather than being explicitly constructed** with step-by-step instructions. For example, a model's ability to perform a task never seen before starting from just a natural language command is an *emergent* property that does not need specific training.

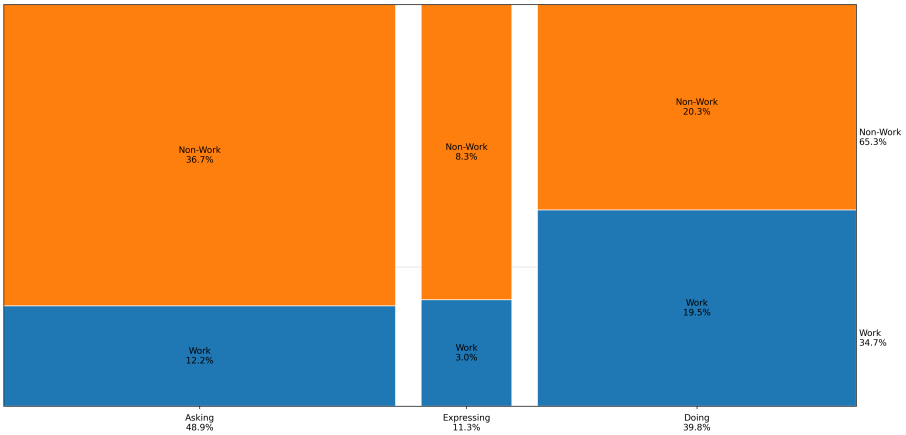
General purpose models are one of the biggest obstacle for the AI Act , as the Act heavily relies on the concept of «intended purpose» [13] , but these models have no specific and defined purpose and lend themselves to a variety of uses, in both personal and professional environments, as official statistics from OpenAI show (see Figure 8.1). The «intended purpose» for a model is indeed very difficult to make explicit once and for all, due to emergence, as history teaches us that a significant share of a model such as GPT's uses were not foreseen neither before it fast appearance on the market, nor after. This fundamental problem remains as of today a topic of contention and cause of uncertainty in the concrete application of the AI Act.

8.4 Innovation *versus* regulation

A long-standing debate is whether the AI Act strikes the right balance between protecting society and fostering innovation, especially when in competition on a global stage against Countries that are much more exploration-oriented. As with other pieces of legislation tasked with going against the unchecked growth of the next technological advancement, there is always a party that is vocally against regulation that could hinder the economic success of entrepreneurial endeavours. Indeed, the EU is already perceived as lagging, attracting only 6% of global AI startup investment in the first half of 2024[72]; this is, at least in part, explicitly attributed to the regulatory burden created by the AI Act, which raise the market entry threshold with compliance costs, and risks stifling good ideas that could pass a more in depth risk-benefit analysis [51], despite authors such as Bradford[16] noting that it was not the laxity of American laws in the first place to enable the leadership role of the USA in the technological ecosystem, so the lack thereof is not the fault of the stricter European digital regulation. Still, as AI is a key area of competition, there is fear of missing out on opportunities that more liberal countries such as the US and China are free to pursue, which are also hindered in their expansion in the European market. Indeed, voices in Europe frame the AI Act and its regulatory tools as an advantage, suggesting that the framing of innovation versus regulation is a «false dichotomy»[117], as industrial policy, including regulation, can be strategically used to create an environment that favours domestic companies. This



(a) Classification in majour generalized work activities of 1.1 million ChatGPT messages.



(b) Shares of messages, over a 1.1 million sample message pool, belonging to Work or non-Work related tasks out, divided by main query goal.

Figure 8.1: Official OpenAI statistics[19] on their models' usage.

is far from a recent debate, as this dichotomy has been challenged widely in recent academic literature and policy analysis, suggesting that the traditional view, particularly advocated by powerful technology firms, that stringent regulation stifles innovation and undermines technological progress, could hide an oversimplification of a much more nuanced relationship. On the one hand, the EU's AI Act, despite industry pressure to dilute it, is argued to **build public trust**, a **prerequisite** for large-scale adoption of trustworthy AI systems; on the other, when regulation is well-designed, predictable, and aligned with market incentives, it can become a powerful engine that compels firms to fundamentally re-engineer their technologies.

The AI Act is often portrayed as a «trust-building» instrument, yet the most immediate obstacle to AI innovation in Europe is perceived as not the Act itself but the way existing data-protection law is applied: Judith Arnal[7] shows that fragmented enforcement across Member States increases regulatory uncertainty and disproportionately affects startups that lack resources to manage varying national interpretations. Without a stronger, binding European Data Protection Board to ensure consistent application of its opinions, the EU risks falling behind in the global AI race.

Indeed, a growing body of research indicates that the relationship between innovation and lax regulation is more nuanced and counterintuitive than it may seem. Regulation can create incentives for firms to re-engineer their technologies to comply with new standards. When successful, these firms regain first-mover advantages and can offset compliance costs through reduced future costs, especially with institutional support such as AI sandboxes. According to the "innovation offset" theory, the benefits of redesigning technology can outweigh the direct costs of meeting new regulations. Further studies by Park et al. [94] on U.S. firms show that regulatory restrictiveness can encourage radical innovation when regulatory uncertainty is low. In these cases, firms use the pressure of tighter rules to search for novel solutions, leading to new patents and technologies. However, when uncertainty is high, this search is stifled, confirming the classic idea that excessive unpredictability can hinder innovation.

Regulatory sandboxes seek to reshape the conflict between regulation and innovation by creating collaborative environments where regulators and innovators work together to co-create future rules. These sandboxes allow for the testing of emerging technologies in a "learning-by-doing" setting, while still maintaining special safeguards to ensure data safety standards are respected.

The so-called *Porter hypothesis* (from Porter and van der Linde[101]) has the opportunity to fully manifest in a new context: that of AI innovation. Introduced in the 1990s, this theory suggested that environmental regulation gives firms a chance to innovate and become more competitive, rather than stifle their growth. Regulation creates pressure to develop better, more sophisticated methods that not only "get the job done" but also uphold higher standards in both function and development. When a company successfully develops new technology to meet regulatory demands, it can gain a first-mover advantage, capturing market share and reaping significant economic rewards. Over time, this theory has been validated by empirical research, especially in the environmental sector. Zhang et al. [127] reviewed 58 studies on the relationship between environmental regulation and green innovation, finding a positive correlation between regulation and

green innovation, though the effect depends on specifics. "Command-and-control rules" (which set explicit performance targets or ban specific activities, such as emission standards) show the strongest and most consistent positive impact on patenting and other innovation metrics, such as patent counts, new product sales, and R&D expenditure. In contrast, market-incentive and voluntary schemes are less robust, sometimes showing insignificant or even negative effects. Developed economies respond more strongly to command-and-control regulation than developing ones, and the relationship is stronger at higher levels of aggregation. In other words, regulatory effects are greater at the country and provincial levels than at the city level[127], suggesting that a larger market size boosts firms' willingness to invest in meeting standards, while for smaller firms, investing in compliance comes less naturally. By extension, EU-wide regulations create stronger innovation incentives than national ones because a firm that re-engineers to meet EU standards can apply that technology across a much larger market. This is a major incentive behind recent developments in EU law regarding the Digital Market, as a unified framework significantly increases firms' motivation to invest in re-engineering and adaptation.

The underlying mechanism is relatively straightforward but powerful: when a regulatory body sets a specific performance threshold—such as an emissions limit, safety standard, or data-protection requirement—firms cannot simply continue with business as usual, since non-compliance carries significant penalties, including fines, restrictions on market access, or loss of license to operate. Faced with this clear choice, firms are compelled to invest in developing new technological components or even entirely new processes to meet the requirements. A firm that develops the most cost-effective, cleanest, or safest solution will gain a competitive advantage over rivals who simply comply at a higher cost through minor adjustments and incremental fixes. In the environmental context, there are many historical examples of regulations sparking entirely new industries: catalytic converters in automotive manufacturing, renewable energy technologies, and pollution control equipment all emerged or expanded significantly in response to regulatory mandates. This dynamic often encourages radical redesigns. The first firm to patent such innovations will enjoy years of exclusive commercial benefits, providing the financial cushion needed to recover their initial R&D investment.

The strength of command-and-control regulation as a driver of innovation lies in its clarity and lack of ambiguity. Firms know exactly what they must achieve, the timeline to follow, and—crucially—that non-compliance is not an option. This alone serves as a powerful motivator to pursue the greater good, even when good conscience falls short. Since compliance has become essential for market participation, these laws effectively align firm profitability with adherence to the rules.

The relationship between innovation and regulation is complex and cannot be reduced to a simple trade-off. Evidence shows that well-designed, consistently applied regulations can promote both market and social innovation. Structural reforms, harmonised implementation, and carefully crafted regulatory tools—such as sandboxes and command-and-control standards—have the potential to align the incentives of firms, regulators, and society. However, a crucial refinement to the Porter hypothesis concerns the role of ***regulatory uncertainty***.

Research[94] on over 1,200 U.S. firms shows that when companies clearly understand

which activities are prohibited and trust that regulations will remain stable, they are more likely to pursue alternative technological solutions with the potential for groundbreaking innovation. In fact, even more radical, disruptive innovation occurs in stricter regulatory environments: empirical evidence shows that destabilising innovation increases by about 1.34 percent of the mean under stable, restrictive conditions, but this effect disappears when uncertainty is high. High regulatory uncertainty discourages firms from seeking new solutions, as innovation is inherently risky and expensive. When firms are already unsure about which technological path to pursue, additional regulatory doubt leads them to delay capital expenditure in R&D, defer hiring, and postpone market entry. The Llama case illustrates this: when the European Data Protection Board took months to decide whether Meta could use public data to train generative AI models, while the UK's data protection authority quickly granted approval [7] through a clear legitimate interest analysis, the difference in regulatory certainty created different incentives. The result was a suppression of the appetite for the kind of exploration that EU legislation is often designed to encourage. This finding reconciles two seemingly contradictory schools of thought in the literature: the resolution lies in recognising that regulatory certainty and the type of innovation are the critical factors in determining whether restrictiveness reduces the range of available technological components, or pushes firms to seek new alternatives. The Park, Wu, and Funk study [94] offers a detailed explanation for why this happens: as restrictiveness increases, some technological components or activities are prohibited, reducing the set of familiar tools firms can use. By itself, this constraint would suppress innovation, especially incremental improvements. However, when restrictiveness is paired with low regulatory uncertainty, the outcome changes. Firms must either accept the constraint or search for alternatives, which is especially costly and risky when future rules are unpredictable. In a stable, predictable environment, firms have confidence that any new components they develop will remain permissible. Knowing the rules will not change, firms see searching for new components as a rational investment rather than a risky gamble. The novel technological combinations they explore become destabilising innovations precisely because they depart significantly from the existing technological architecture. Firms are not simply complying; they are actively re-engineering their technological approaches to find new solutions in response to the incentives created by clear, binding restrictions and regulatory stability.

8.4.1 The structural context as a moderating factor

While regulatory incentives can be potent, their true effectiveness is contingent upon the broader structural context in which they operate, as a firm's ability to respond to these incentives is shaped not solely by the regulatory framework itself, but also by such factors as access to capital, the prevailing legal environment governing bankruptcy and risk, overall market size, and the availability of skilled talent.

The United States' technological leadership in the digital sector cannot simply be credited to a permissive regulatory climate, as is sometimes claimed. Instead, it stems from a combination of **structural strengths**[16]: a unified digital market that enables agile scaling, robust venture capital funding for high-risk research and development, bankruptcy laws that give entrepreneurs a second chance, and immigration policies that continually renew the talent pool. By contrast, in the European Union, even the best-

designed *command-and-control* regulations do little to overcome these underlying cultural differences, which inherently limit firms' ability to innovate as quickly as their U.S. counterparts.

None of this makes regulation irrelevant, but it does mean that regulation must be part of a broader policy toolkit to be truly effective. For regulation to act as a genuine catalyst for innovation, it must be implemented alongside structural reforms that give firms the financial and organisational resources they need to fully realise their capabilities.

Academic reviews[16, 94, 127, 104] on innovation and regulation highlight several strategies that reinforce each other for policymakers aiming to use regulation to foster innovation. First, guarantee regulatory certainty by setting clear, stable rules (as Arnal[7] points out, *«[it is] The implementation of the General Data Protection Regulation (GDPR) in the EU, rather than the EU's regulation itself, is holding back technological innovation.»*), establish **binding authority for oversight bodies** (such as the European Data Protection Board), publish detailed compliance guidance, and commit to predictable regulatory baselines; when firms lack clarity or confidence in the regulatory environment, they tend to default to cautious, non-innovative behaviours. Second, blend *command-and-control* standards with market-based incentives, so all firms meet a solid performance baseline while being encouraged to exceed it for extra financial or reputational rewards; this hybrid approach combines regulatory certainty with market dynamism. Third, create regulatory sandboxes and collaborative spaces with safeguards against regulatory capture, such as transparent entry standards, public progress reporting, independent evaluation, and strict rules against arrangements that unfairly benefit incumbents. Finally, fix structural gaps that limit firms' ability to respond to innovation incentives by completing the Digital Single Market, broadening access to venture capital, and adopting immigration policies that attract and keep top talent.

According to the cited authors, these systemic changes are crucial for unlocking the innovative potential that regulation can bring. When used well, these tools let organisations pursue bolder goals than traditional regulatory processes would allow, boosting their competitiveness.

8.4.2 Trust-building and the market-access pathway

In addition to *command-and-control* type legislation, regulation can promote technological development by fostering trust and expanding market access. The European Union's proposed AI Act famously champions this approach. The EU Commission maintains that the regulation enhances innovation by building the trust required for widespread AI adoption. When citizens recognise that AI systems are subject to robust safeguards, their willingness to accept and utilise these technologies increases. This dynamic establishes a positive feedback loop: uniform EU-wide standards reduce regulatory fragmentation across multiple states, rather than requiring adaptation to distinct national regimes. As compliance with the AI Act's requirements signals a high level of production standards, the mere act of compliance can signify competitive advantage, potentially commanding premium pricing, and can reframe the re-engineering required to meet those standards as a value-creation opportunity rather than a mere cost.

8.4.3 Market-incentive regulation and voluntary schemes: weaker but complementary mechanisms

Market-incentive mechanisms, like subsidies and tax breaks, work differently from *command-and-control* standards and create distinct incentive effects. Instead of requiring a specific outcome, these mechanisms reward companies financially for exceeding the minimum required performance. In theory, this should strongly encourage re-engineering, which might be more appealing than the threat of a fine—as firms that invest in breakthrough technologies can reap financial rewards for going beyond the baseline. However, as mentioned earlier, empirical evidence[127] shows that *market-incentive* regulation leads to weaker and less consistent innovation than command-and-control approaches. Zhang et al. suggest this is because market incentives set a ceiling for compliance, not a floor for further progress: companies focus on securing subsidies or tax breaks rather than pushing for more radical changes that fall outside current funding calls. Worse still, if the financial incentive is too small compared to the investment required, firms may simply ignore it and drop plans to improve.

Voluntary regulation, such as industry pledges to follow best practices or *self-labeling*, usually provides even less motivation for innovation than other methods. Since self-assessments and voluntary schemes are generally less strict and more subjective, they demand little real compliance and make it easy for companies to participate. Zhang et al.'s research shows that while these programs might lead to slight bumps in patent filings for compliance, they don't result in higher sales—indicating small tweaks rather than the significant innovations needed to succeed in the market.

Despite these drawbacks, *market-incentive* and voluntary approaches can still support *command-and-control* regulation. If a regulatory system sets a firm baseline that all companies must meet, but also offers extra rewards or recognition for those that go beyond, it can take advantage of both strategies: the baseline ensures necessary improvements, while the incentives encourage more companies to outperform the minimum requirements.

8.4.4 Regulatory sandboxes as collaborative redesign spaces

Articles 3(55) and 57 of the AI Act require European Union member states to set up at least one national AI regulatory sandbox. These sandboxes are supervised spaces where high-risk AI applications can be developed, trained, tested, and evaluated for a limited time before entering the market. Their main goal is to support technological progress, especially for SMEs and start-ups, by offering flexibility and shielding them from administrative fines under the AI Act (and other EU or national laws that the relevant authority[8] is competent in) as long as they follow official guidelines in good faith. In these settings, participants receive regulatory guidance and a secure environment to proactively tackle risks—particularly those related to fundamental rights, public health, and safety.

[8]The supervisory framework of the AI Act distinguishes between several national competent authorities. Market-surveillance authorities are often seen[49] as the most suitable operators because they already monitor AI systems after they enter the market, but also notifying authorities may also act as competent authorities. In Italy the Sandbox program is managed by the Department for Digital Transformation within the Prime Minister's office, with the Agenzia per l'Italia Digitale (AgID) providing technical support and evaluation during experimentation[63].

Developers also get clear direction on compliance requirements, and completing the sandbox process demonstrates their readiness for the market. Another key benefit is the chance to safely use personal data for projects in the public interest, as long as they provide thorough documentation of their activities and risk management—all mostly free of charge.

Regulatory sandboxes provide a different way to use legal tools to drive innovation. In these controlled labs, regulators and innovators work together to test new technologies and create temporary standards that can shape future laws, much like the process in IT standard-setting. A recent ethnographic study [69] of a UK healthcare regulatory sandbox for AI in diagnostics illustrates this approach, as it highlighted how the sandbox fostered innovation through three practices: **protection, adaptation and participation**.

Protection: a sandbox kept a separate environment using confidentiality agreements and an *ad-hoc* approach, reducing firms' concerns about exposing vulnerabilities or facing enforcement.

Adaptation: the sandbox supported the iterative development of regulation, with **regulators and regulatees collaborating** to develop standards through feedback and iterative refinement. This adaptive process encouraged firms to feel engaged and motivated in the endeavour of technological redesign to meet law requirements.

Participation: the sandbox brought together diverse stakeholders, including healthcare providers, technology suppliers, patient representatives, clinical experts, and regulators. This brought together many different point of views on trustworthiness, safety, and ethics, not just technical compliance; for example, the involvement of patient advocates prompted developers to address the actual lived experiences of the subjects, not just the machines' purely technical standards.

As a sort of "liminal space," sandboxes encourage firms to move away from their established business models, collaborate to define best practices, and reshape their strategies and paths: instead of seeing regulation as a distant constraint that «fell from above», participants become co-creators of the regulatory framework, and have a change to engage in a kind of technological re-engineering that is strictly interwoven with the goals of the regulations.

8.4.4.1 The risk of regulatory capture in collaborative spaces

Regulatory sandboxes and other collaborative regulatory approaches can spark innovation, but they also come with a major risk: *regulatory capture*[104]; this happens when regulators and the organisations they oversee spend too much time in close contact—especially if participation is selective and the process lacks transparency—which can let well-resourced firms shape new standards to suit themselves, often to the detriment of smaller competitors. Ranchordás and Vinci[104] highlight Italy's Sperimentazione Italia as an example: despite aiming to promote responsible innovation across sectors, the sandbox received few applications and generated minimal collective learning[104], largely because of a burdensome application process and a sense of unpredictability and

bias. Ultimately, the promise of faster market access failed to attract firms without the resources and reputation of established players, firms that could afford the risk of rejection. Another issue was the lack of publicly available information about project approvals. While it's reasonable to protect confidential business data, this level of opacity made it harder to hold regulators accountable and discouraged potential participants. To guard against capture and make sure sandboxes actually foster innovation (not just entrench the power of big players), it's crucial to put in place strong transparency in candidate selection, clear participation rules, independent evaluations of outcomes, and clear routes for scaling successful projects beyond the sandbox. Without these protections, collaborative regulatory efforts risk becoming vehicles for corporate *rent-seeking*[9], as, according to Ranchordàs and Vinci, it creates scenarios where regulators are **systematically exposed to arguments** from firms that may not align with the public interest.

To reduce the risk of regulatory capture and genuinely encourage innovation free from the influence of dominant players, robust safeguards are essential. Transparent admission standards, public reporting of progress and results, independent evaluations, and strict bans on arrangements that favour certain incumbents are key. Good governance means more than just procedural transparency: it also requires protection from traditional regulatory pressures, ongoing adaptation, and the involvement of a broad range of stakeholders. When these practices are put in place, as seen in the healthcare AI sandbox studied by Macrae and Ansell[69], they can shift the relationship between regulators and firms from adversarial and top-down to collaborative, while upholding legal standards. Without these protections, sandboxes risk making regulatory fragmentation and market inequality worse, undermining fairness and the possibility of broad, generalisable benefits. Ranchordás and Vinci ultimately argue that regulation must always **prioritise the protection of fundamental rights** over the pursuit of technological progress or narrow economic innovation.

[9] *«Rent seeking, a concept introduced by economist Gordon Tullock, refers to the pursuit of wealth without contributing to society's overall productivity or well-being.»[70]*

Chapter 9

Supervision and enforcement

9.1 Relevant authorities

The effective enforcement of the Regulation relies on significant resources (infrastructural, technical, financial and human[1]) and the impartiality of surveillance authorities, which may vary considerably case by case. To ensure harmonisation and effective oversight in all Member States equally, the AIA **mixes Member State-based and centralised enforcement**, requiring the creation of several new **supranational** regulatory bodies (model that has found increasing adoption in EU law in recent years[115]). While most enforcement duties, such as monitoring, investigation and sanctioning, are entrusted to national bodies, some particular topics, such as General Purpose AI models, are handled by the *AI Office* of the European Commission, an *AI Board* composed of representatives from Member States, and *Scientific Panels* of independent experts when in need of additional advice.

9.2 Enforcement challenges

Effective enforcement of the Act ultimately depends on the resources and technical know-how of national Authorities [76] and the effectiveness of the newly established European *AI Office* in guiding the Act's application[81]. Traditionally, the EU has relied on passing large volumes of legislation, yet recent political debates have called for **cutting bureaucracy to boost economic growth**, especially by reducing reporting requirements, in favour of more subsidies and IPCEI [2] projects. This change is driven by concerns that excessive regulation has **held back the digital sector**, but critics warn that removing rules might only yield surface-level improvements. Moreover, inconsistent enforcement of *current* rules can hurt AI innovation even more, because it leads to uncertainty, higher compliance costs for start-ups, and less investment. Only by taking a nuanced, context-aware approach, policymakers can turn regulation into a genuine driver of sustainable, trustworthy innovation.

[1] Including, per Art. 70 of the AIA, «*personnel permanently available whose competences and expertise shall include an in-depth understanding of AI technologies, data and data computing, personal data protection, cybersecurity, fundamental rights, health and safety risks and knowledge of existing standards and legal requirements*».

[2] Member-state funded *Important Projects of Common European Interest*.

9.3 Recent conflicts between Big Tech and European Law

Regulatory actions by European regulatory bodies concerning data protection, AI usage and digital competition[3] have created significant conflicts with major AI service and model providers, leading to a reactionary response to what was felt as a significant threat to the world AI market.

The most prominent clashes actually stemmed from the enforcement of the GDPR, as national *Data Protection Authorities* (DPAs) have been questioning the security, transparency and nature of data (personal or even sensitive in nature) used to train the AI models. For example, a key topic of contention has been which legal basis, if any, can excuse the tech companies' pillaging of personal data available on the Internet.In this fight, incidentally, Data Protection Authorities are joined by media publishers such as the *New York Times*, which famously **sued OpenAI for training their models on copyrighted content scraped from their website** [113], while The Associated Press has settled the issue by formally licensing [6] their text archive (which probably would have been used to train models anyway, for free).

Over time, regulatory authorities have made varied inquiries on AI models, with motivations ranging from the timing of publication of Google's Palm's DPIA[4][26], to the lawfulness of web scraping and the possibility for data subjects to exercise their GDPR-given rights (access, rectification, and erasure)[27].

In front of the uncertainty behind their methods and conduct, DPAs have also been known to take direct action before waiting for answers, forcing major technology companies like Google, Meta, X, and LinkedIn to pause or delay AI projects within the EU[98], as it was case regarding the Italian *Garante per la Protezione dei Dati Personali, Deepseek* [97] and ChatGPT[96] (which was briefly banned in the country in March 2023, citing «*no legal basis underpinning the massive collection and processing of personal data in order to 'train' the algorithms on which the platform relies*» and hallucinations). This particular story continued even after the original frictions, as **OpenAI** was issued a **fine** in December 2024 for GDPR non-compliance[22].

Ireland's Data Protection Commission (which is the governing authority of many tech giants due to them having their European establishment in the Republic of Ireland) also acted with uncharacteristic[120] speed (for one of the European Countries that has been catering the most to Big Tech in an effort to attract huge investments) in sanctioning big names such as.One of the most significant cases in which the Irish DPC was involved was the **blocking of the use of public user data from EU Facebook and Instagram accounts for model training**, following a **complaint** [102] lodged by the privacy rights organisation *Noyb* (*None of Your Business*), founded by **Max Schrems**. The decision represented a «*U-turn*» by the Irish authority, which **had initially approved Meta's AI applications**. Following intensive engagement with the DPC and pressure from its fellow EU DPAs, Meta officially announced its decision to **pause its training plans**. Noyb suggested that this outcome was a direct result of the pressure mounted by them and other like-minded organisations, specifically citing eleven other complaints

[3] See: the EU Digital Markets Act.

[4] Data Protection Impact Assessment.

filed with various national DPAs in Europe. Meta is still withholding the licensing of the multimodal (i.e. capable of processing video, audio, images, and text) version of its Llama model (and its derivatives by other publishers) from European-based enterprises [75], citing too much regulatory uncertainty over training models on data from European users. This outcome mirrored similar moves by other US tech giants, such as Apple and Google, which similarly delayed or withheld new AI features in the EU due to concerns over compliance with GDPR, the AI Act, and the Digital Markets Act: in 2024, Apple publicly stated it would not release some Apple Intelligence features in the EU due to the DMA's **interoperability requirements**, which Apple claims would jeopardise user privacy and data security [20].

Noyb's complaint on *hallucinations* In April 2024, Noyb lodged another official complaint [91] with the Austrian Data Protection Authority against OpenAI, focusing on the tendency of Large Language Models to generate so-called «hallucinations», that is, plausible but nonetheless false outputs. The issue was exemplified by the model's provision of an incorrect birth date for a public figure, a mistake which Noyb argued amounted to a breach of the GDPR's **accuracy principle** (*Article 5(1)(d)*). In response to these allegations, OpenAI reportedly argued that, **due to the inherent technical limitations of LLMs, it was impossible to guarantee the right to rectification** for cases such as these; instead, it proposed **filtering or blocking** inaccurate data **when prompted by specific user complaints**. Noyb, objected to this approach, insisting that OpenAI could not selectively decide which parts of the GDPR to honour, and that the right to rectification is non-negotiable and must be fully respected. Furthermore, Noyb alleged that OpenAI had failed to comply with the **data subject's right of access** as outlined in *Article 15* of the GDPR, because the company did not provide sufficient disclosure regarding what personal data was being processed, the sources of such data, or the exact recipients to whom it might be disclosed [114].

The ChatGPT Taskforce The EDPB's so-called *ChatGPT Taskforce* has been the supra-national referent to reach better coordination in issues such as these regarding the popular AI Model.Initially, the task force was also needed to compensate for the impossibility to apply the *one-stop-shop* rule for OpenAI, which had not yet set an establishment in the EU. Their work culminated in the publication of a **report**[37] on May 23, 2024, detailing the preliminary findings on the adherence of ChatGPT to the General Data Protection Regulation (GDPR)[5]. The EDPB Taskforce report established critical, preliminary compliance standards for Large Language Models (LLMs) under the GDPR, providing *de facto* guidance for Data Protection Authorities before the AI Act came into existence, and setting a clear direction for more investigations. Most interestingly, this report was instrumental in setting clear boundaries against Big Tech's usage of the «*technical impossibility*» excuse as justification for imperfect compliance with the GDPR, and set a precedent for how DPAs should handle AI companies, namely with strict investigations and audits on transparency, data security, accuracy, and the concrete

[5] More on the friction between the GDPR and the AI world is presented in 12.

possibility to exercise the rights given by European legislation.

While firms like Amazon, Microsoft and OpenAI sign up for a voluntary commitment to the principles of the AI Act (the AI Pact [90]), other big names such as Meta and Apple keep their silence as a symbol of their long-standing fight [34] to remove or at least simplify their path to (formal) compliance. As AI providers keep clashing with the EU over tight rules coupled with heavy regulatory uncertainty, delayed decisions, conservative interpretations from the EDPB and the unpredictable environment that is created, major technology companies keep launching significant lobbying efforts, together with a long list of European stakeholders (mainly in e-commerce and data science). A famous open letter [33] by them argued that the EU's approach risks AI models that «*won't understand or reflect European knowledge, culture or languages*», and dangled before European legislators the «carrot» of billions of Euros in investments (with a consequent increase of more than 1.2 trillion Euros to the EU's GDP over ten years[98]). Even the strategy of voluntarily withholding products from the market can be seen as **another way of putting pressure on European Union policymakers** in this sense [125]. In a way, the delays and cancellations uncover this double mask of Big Tech in Europe: one day allies and benevolent suppliers of wonderful technical instruments , and the other determined *colonisers* and rivals in the race for (much more than) market dominance.

9.4 The current state of Transparency in the industry

The *2024 Foundation Model Transparency Index* compiled by Bommasani et al. [14] assessed 14 foundation model developers against 100 transparency indicators. The average score for developers was 58 out of 100, with the highest of 85 (more than one standard deviation above the mean) reached by the BigCode, **Hugging Face**, and ServiceNow developers, and the lowest , being more than one standard deviation below the mean, were **Amazon** and Adept.

As expected, the results show that developers are generally most transparent about **downstream aspects**, such as usage policies, and least transparent about the **upstream resources** used to build their models, i.e. data and algorithms. Breaking the domains down further reveals specific areas of transparency and opacity, with the least transparent subdomains being **data access policies**, **impact** assessment, **trustworthiness**, and **mitigation** techniques.

The index also compares developers who release model weights openly (6 developers) with those that use a more closed, API-based strategy (8 developers). Overall, *Open* developers generally outperform closed developers in transparency, especially in the upstream domain. Closed developers only preferably share information regarding topics such as policy enforcement, risks, and model mitigations.

As this was the second round of inquiry made with the same framework, it was possible to record the change in just a year's difference: despite overall improvements, **transparency** on data access **declined** from 20% to 7%, and information about data copyright status, data creators, and personal information did not improve.

9.4.0.1 Google and OpenAI

It is worth it to spend a bit more time on the results concerning two of the biggest AI names in Europe at the moment: **Google** and **OpenAI** (GPT family models).

Out of its 49 overall points, OpenAI's transparency was weakest in the upstream domain, which covers the resources used to build the model, but it did not fare well in any of the calculations in general. More specifically, it scored 0 on indicators related to *data creators*[6], *data copyright* [7] *and license status* [8], *direct data access* [9], and broader *environmental impact*[10]. It also scored 0 on all indicators for *trustworthiness evaluation* [11]and *inference evaluation* [12], as well as on third-party evaluations for the model's **limitations** and **risks**, and on **real-world impact**, getting a 0 also on indicators for *affected market sectors* [13], *affected individuals* [14], *usage reports* [15], and *geographic statistics* [16].

Google was instead assessed based on its flagship model, **Gemini** 1.0 Ultra (only accessible via API). Its overall score of 57 was the reflection of a situation similar to OpenAI's, despite it having a better score in the *model* and *downstream* domains: as OpenAI, Google scored 0 on indicators for *data creators*, *data copyright* and *license status*, *direct data access*, and broader *environmental impact*; on the other hand, it demonstrated greater transparency than many peers in **risk and mitigation evaluation**: it was one of only three developers to score points for **trustworthiness evaluation** and one of two for **external reproducibility**. It also scored well on both unintentional and intentional **harm evaluations**. The same cannot be said for real-world impact evaluation, which scored 0 on the same four impact indicators as OpenAI (*affected sectors*, *individuals*, *usage reports*, *geographic statistics*).

9.4.1 No transparency = no trust = no trustworthy results

Warren J. von Eschenbach's paper[32] on the issue of black box systems shows from a philosophical perspective just how radical and profound is **the conflict between the need for trustworthy AI and the current panorama of closed-source models**, as he argues that the opacity of these systems **prevents them from meeting the**

[6]Demographic or identity information about the human authors, annotators, or content producers who created the training data.

[7]Indication of what portion of their training data is protected by copyright.

[8]Licenses under which the data was obtained

[9]The existence of a mechanism for external parties, such as researchers or auditors, to directly access and examine the training dataset.

[10]Information about the environmental costs of their model beyond direct carbon emissions from training

[11]Whether the developer conducts and discloses the results of formal evaluations of their model's trustworthiness.

[12]Costs (energy and computational resources) and performance (latency from request to answer) of the model when being used to generate outputs.

[13]Whether a developer identifies and reports on the specific industries or economic sectors that are most affected by their model's deployment.

[14]Information about the specific demographic groups, professions, or types of individuals most impacted by the model's use in the real world.

[15]Aggregated, anonymised statistics about how their model is being used.

[16]Data on where in the world their model is being most heavily used

necessary conditions for trustworthiness, even when they are highly *reliable* in describing factual truth.

The heart of von Eschenbach's discourse is the fundamental distinction between *trustworthiness and reliability* : the former is a *moral* quality that is not reached by just knowing that the model will be competent in performing a task. That, he says, will be enough only to justify *reliance*, which is the act of **delegating** a task and expecting it to be carried out correctly. The concept of ***Trust***, on the other hand, involves the delicate act of **permitting oneself to be vulnerable towards an actor** (the AI model), based on the belief that the trustee **will act in the trustor's best interests** (he difference becomes clearer when we think that no one feels betrayed when a computer crashes, but we very much feel so if a good *friend* we *trust* hurts us).Trusting something, von Eschenbach says, means the output of the delegated task is not the only important part: **the motivations and methods used to carry out the task must align with the trustor's values and interests to establish a true trust relationship**. Since black-box models do not show the values and inferences that underlie their decisions, it is still not possible to establish a truly trustworthy AI. On the contrary, it could be said that AI systems are most certainly shaped by their *owners*' goals, which may or may not coincide with the public good. This is a particularly important point to make as we enter the realm of **high-stakes decisions** across criminal justice, finance, and healthcare. In those fields, the importance of understanding *why* a conclusion has been reached remains paramount to enable humans to confidently act on it and eventually *learn* from its insights without becoming subordinates and losing control of the outcomes. Until now, these contexts have preferred what is called *social* vetting, where **the human remains the *trusted* party**. Von Eschenbach's thesis hinges on exploiting this delicate distinction by going from the question of whether to trust a *model* to whether to trust the socio-technical system that has included and accepted AI into its operations. For most people, he predicts, the relationship with AI will remain **mediated** by the experts and institutions responsible for its deployment, which in turn trust each other in doing their due diligence in their roles; for example, the patient will trust **the doctor's *interpretation*** of an AI-augmented diagnosis, while the doctor will trust both the *developers* and the *regulatory bodies* that have built and thoroughly tested the tool. This point of view also addresses the age-old view of AI as a tool less influenced by human emotions and incentives than human decision-makers: by making them both keep each other in check, perhaps a new kind of balance can indeed be reached.

Von Eschenbach's view sees legitimation in the words of another author: Marco Almada of the University of Luxembourg. He writes, in his paper titled *Technical AI Transparency: A Legal View of the Black Box* [5], that **what the law demands from transparency is more complex than what technical approaches alone can deliver**, because the requirements are rooted in broad, long-lived principles (like the European Convention of Human Rights , the constitutions of individual member states, and pre-existing horizontal legislation such as the GDPR), each of those serving purposes beyond mere visibility of a system's mechanics. From this point of view, the demand for AI transparency is not even a new concern but an extension of an existing approach to new technologies, one that is technology-neutral and focused on setting fundamental

principles that stand the test of time. This approach delegates the responsibility for specifying the precise technical details to courts and administrative bodies, who must interpret these general principles on a case-by-case basis.

Almada's paper proposes an holistic, multi-faceted strategy for learning to «*liv[e] with the black box*» rather than abandoning transparency as an ideal; this includes to expand **disclosure requirements** beyond technical artefacts to include key design choices and information, a stronger reliance on third party auditing, and, in cases where the impact to individual rights threatens to be high, even prohibit the existence of certain too-opaque models, as the European Court of Justice did when it found that self-learning systems were incompatible with the rules for processing passenger name records [67].

9.4.2 Justifications from the industry

The issue of transparency is often described as a conflict between commercial interests—specifically, the desire to protect proprietary innovations from market incumbents—and the public's need for external validation and safety assessment. Sapkota et al. [111] highlight in their research how detailed documentation regarding the origin, composition, and cleaning of training datasets **is routinely missing for almost all the main models on the market** (including Meta's Llama, Deepseek, Google's Gemini, BERT and T5, Med-PaLM M, Mistral 7B, etc., with the notable exception of Dolly 2.0 [17]).

9.4.3 On-the-field analysis on transparency and accessibility

The February 2025 *Comprehensive Analysis of Transparency and Accessibility of ChatGPT, DeepSeek, and other state-of-the-art Large Language Models* [111], examines how transparent and accessible some widely discussed models really are. The report highlights a gap between the way some models are marketed as «**open-source**» and their true nature, which is better described as «open-weight». This is the case, for example, of Meta's **LLaMA** series and **DeepSeek-R1**, which keep their datasets proprietary making full, independent reproducibility impossible. Proprietary control often extends to key components such as fine-tuning methods(and reinforcement learning RLHF pipelines, and hyperparameter schedules) and complete implementation details. including, code and infrastructure (such as including cluster configurations and compiler optimisations), and architectural details (like mixture-of-experts (MoE) routing mechanisms). Notable examples of missing information include:

- The *reinforcement learning from human feedback* **pipelines** for models like **DeepSeek-R1** and **Mistral** AI.
  - Reinforcement Learning from Human Feedback (**RLHF**) plays an essential role in guiding intelligent systems to align more closely with human preferences and intentions. This approach allows machines to learn effectively from observing human behaviour to adapt their policies . From RLHF, the model can derive an understanding of the underlying reward functions that

[17] Dolly 2.0 is described by its developers as *«the first open source, instruction-following LLM, fine-tuned on a human-generated instruction dataset licensed for research and commercial use»*[25].

the expert is maximising, aligning its behaviour to the one that is rewarded by the trainer. [109]

- Hyperparameter schedules and learning rates , such as **Microsoft Phi-4**'s and **Google Gemini** Ultra's.
  - **Hyperparameter schedules** are fundamental aspects of optimising machine learning models, impacting their efficiency and effectiveness during the training phase. **Hyperparameters** are crucial as they govern the learning process and model performance, but are not learned from the training data itself; instead, they are set prior to training. One of the key hyperparameters is the *learning rate*, which determines the size of the steps the optimiser takes during the training to minimise the loss function. The «*schedules*» are the predefined strategies for adjusting hyperparameters during the training process.
  - The **learning rate** plays a significant role in controlling how quickly a model adapts to the problem it is trying to solve, and, ultimately, the model's performance: if the learning rate is set too high, the model may converge too quickly to a suboptimal solution or diverge altogether. Conversely, if the learning rate is too low, the training process may become very slow and could get stuck in local minima, thus failing to reach an optimal solution. [109]
- Either weights, training code, or deployment environments for 68% of the analysed models.
  - **Weights**, **training code**, and **deployment environments** collectively dictate how an AI model operates and how accurately it delivers output. They shape the model's learning process, generalisation abilities, and performance in real-world scenarios. The calibration of **weights** determines how the model generalises from the training data to unseen examples, the **training code** defines how data flows through the model and how learning occurs, affecting its ability to minimise the loss and, consequently, predict accurately, while the way the **deployment environment** is managed impacts how quickly updates can be rolled out, which is crucial for maintaining model relevance in a changing context. [109]
- Google Gemini's multimodal alignment loss functions for its *cross-modal fusion logic* and the specific *MoE router logic* for its mixture-of-experts (MoE) architecture.
  - **Multimodal Alignment Loss Functions** ensure that different types of data (e.g., text, images, audio, videos) can be used together to make predictions based on a richer context. This alignment is fundamental to the model's ability to interpret information from multiple angles and provide **relevant** outputs. [109]
  - **MoE Router Logic** directs inputs to specific *experts* in the model, which are **specialised sub-networks** trained to handle particular tasks. This routing

mechanism promotes higher efficiency in how computational resources are used. [109]

All in all, the investigation reveals a **consistent pattern of selective transparency**, from the same companies that, when convenient, promote community engagement through forums, user guides, and documentation to encourage wider adoption. Even if AI developers such as DeepSeek have indeed been very interested in setting strategies to commoditise foundational AI, significantly lowering usage costs and making powerful tools more accessible to smaller businesses and individual developers, the authors argue that this is not sufficient to make up for the lack of comprehensive transparency needed to fully understand internal mechanisms. Restricted access to fundamental data and methodologies systematically impedes the replication of research findings, third-party error analysis, and the identification of inherent biases within models, doing nothing to counteract ongoing concerns regarding the reliability and trustworthiness of LLMs in critical applications.

9.4.4 A proposed solution: shifting to open-source models

Although they have been unfairly overlooked by the global mainstream due to perceived lower performance, open-source models offer compelling opportunities for reduced compliance burdens and a more transparent, equitable approach—especially when used in sensitive, high-stakes public applications.

The decision to «close the source» of an AI Model is dictated by strategic motivations to prosper in the technology market, and the great amount of money that these companies have obtained from being at the top, scarcely undisputed, has allowed, in return, for even bigger investments in the next high-performing model. Without the same promises of monopolistic control to the fruits of the researchers' work, and with fewer opportunities to capitalise on their development, open source models have been neglected to the point that it is still largely believed that they remain less performing and of lower quality than their closed counterparts. This, however, has been slowly becoming falser, as studies found that leading open-source models had indeed caught up to and, in some cases, surpassed GPT-3.5-turbo (ChatGPT) on various benchmarks[21]. In 2023, this represented a major milestone, demonstrating that the open-source community had successfully replicated much of the capability that initially seemed exclusive to well-resourced tech giants. Among those analysed, Llama-2 by Meta and the Technology Innovation Institute's Falcon were considered the leading open source models, although the former has not been recognised[126] by the OSI standards organisation as fitting their definition (see section 9.4.4.1) due to transparency and licensing issues, while the latter was deemed adequate if not for its restrictive licence. Among the models that were recognised as fully compliant, i.e. Pythia (Eleuther AI), OLMo (AI2), Amber & CrystalCoder (LLM360) and T5 (Google)[82], OLMo 2 7B is the one that has attracted more attention for comparisons, but only with Mistral 7B, outperforming it in mathematical problem solving and multitask language understanding, while lagging behind in code generation [118].

Recently, OpenAI has released a new open-weight [86] model as well, *gpt-oss*, under what is described by the company itself as a «permissive» [85] Apache 2.0 licence.

True openness, however, requires more than just freedom, but **access** to information on how the model was trained. In OpenAI's own words [84], these models «were trained using [...] techniques informed by OpenAI's most advanced internal models, including o3 and other frontier systems.» Additionally, while advertising greater control and flexibility given by the choice of gpt-oss over closed models, OpenAI stated that «While the trained weights are open, some surrounding infrastructure or tooling may remain proprietary to their providers.» [87]; as *tools* in this context signify various fundamental components such as training, validation and evaluation code [124], this could very well mean the end of their road for true open source, as what is systematically withheld is most valuable and sensitive assets: the comprehensive code that defines the training process, the full composition of the training data, and the specific methodologies used for crucial stages such as reinforcement learning from human feedback.

9.4.4.1 What is Open Source AI

True open-source models, as defined by standards like those set by the Open Source Initiative (OSI), demand unfettered access to the entire stack: the model architecture, the complete training code, the training data composition, and the resultant model weights. Furthermore, they require disclosure of sustainability metrics, such as carbon emissions, and explicit details concerning bias mitigation strategies. The goal is complete clarity, enabling researchers and regulators to replicate and critically examine the system's behaviour.

As defined by the Open Source Initiative (OSI)[83], a true open source AI system is «*made available under terms and in a way that grant the freedoms to:*»

> **Use** the system for **any** purpose and without having to ask for permission.
>
> **Study** how the system works and inspect its components.
>
> **Modify** the system for any purpose, including to change its output.
>
> **Share** the system for others to use with or without modifications, for any purpose.

Many prominent Large Language Models that claim to be open source often fall short of the traditional Open Source definition because of licensing restrictions. Traditional open-source licenses prohibit restrictions on specific users or fields of use; as a result, these models are now typically classified as "open-weight" rather than truly open source. A notable example is Meta's Llama series: these models are licensed under a custom Llama Community License Agreement, which specifically prohibits use by entities with over 700 million monthly active users and requires agreement to its Acceptable Use Policy. While these policies are designed to reduce the risks associated with improper use of AI models, they effectively disqualify the models from being considered truly open source. Critics of truly open source AI [93] argue that there is a high risk of misuse, and that powerful tools could end up in untrustworthy hands without the ability to control access or implement built-in safeguards, as is possible with closed-source alternatives. On the other hand, open-source solutions offer greater transparency and control over data; for example, having access to free, downloadable weights allows users to run the model on their own infrastructure, avoiding the need to send queries to remote APIs

and simplifying compliance with data protection regulations. Furthermore, the ability to fine-tune the model enables the creation of specialised tools tailored to the user's needs [88].

An AI system is more than just its license agreements; it consists of three key components:

- the dataset(s) and processes used to train it,
- the architecture and code that defines and shapes it, including those scripts used to process the input datasets and the inference code, used to run the final model and give new outputs.
- and the model weights, meaning the numeric parameters that represent what the model has extracted and «learnt» from the training data [62].

Sapkota et al. [111] identify as key metrics for true open-sourceness the following pieces of information that are needed to really evaluate a model in in the qualitative and quantitative realms:

1. Licensing, Usage, and Redistribution Rights
2. Code Accessibility and Modification Rights
3. Training Data Transparency
4. Community and Support
5. MMLU Score and Carbon Emissions
6. Ethical Considerations and Reproducibility

In today's AI development landscape, the term "open" is often used as a synonym for "not-so-closed," a phenomenon commonly referred to as "open-washing." While developers may market their models as open source, a closer look reveals that many leading LLMs only meet the criteria for being open-weight models. Open-weight models are currently the most popular compromise between open source and total proprietary control: their training parameters (weights) are publicly accessible and available for inspection, but other information such as training data and codes remain secret, and usage is legally restricted, especially for commercial enterprises: models such as Llama require a separate commercial licensing agreement with Meta once a certain usage benchmark is reached. Similarly, Mistral Large requires direct commercial agreements for production-scale deployment to allow the original developers to retain control over large-scale commercial exploitation.

According to Kapoor et al. [60], open-weight models have **five distinctive properties** that bring significant societal benefits as well as potential risks. While the benefits are substantial in theory, the risks of misuse are often poorly understood, especially with regards to **the marginal risk**, that is, the additional risk these models pose compared to closed models. Closed-source models are typically accessed only through tightly controlled APIs, but open-weight models are characterised by the following unique properties[60]:

- Broader **access**: The model weights, an important part of how they reason, are widely available.

- Greater **customisability**: Users can modify the model through fine-tuning for specific applications.
- Ability to run the model **locally**, on local hardware.
- Inability to rescind **access**: Once the weights are released, they can be copied and redistributed indefinitely.
- Weaker **monitoring**: The developer has little to no ability to monitor how the model is being used (especially when run locally).

Kapoor et al. identify from these some **key consequences** of choosing open weight over closed source models, as the former allow for:

- Distribution of **power**: Open models prevent a handful of large tech companies from having unilateral control over what AI can and cannot do. Broader access and customisability allow more people to explore and build new applications, especially in sensitive areas where using a third-party API is not advisable. This enhances privacy and security for applications involving sensitive data and encourages entrepreneurs and startups to develop their own original ideas without huge upfront investments.
- Loosely-added safety features implemented by the original developer can be removed via customisation processes.
- No control on who copies and uses the weights means the developer cannot take them back, even if they find them to be faulty or biased. For all intents ans purposes, releasing the weights is irreversible and fundamentally shifts control from the developer to the user, making monitoring and moderation nearly impossible for the original creator.
- Transparency: access to weights give much more data to scientific research on reproducibility and safety, which is basically impossible to conduct independently with closed models, to search for bias, security flaws, and other faults.
- Mitigation of the **monoculture** effect: By allowing for greater diversity in downstream models, openness reduces the systemic risk of having society depend on a small number of models with shared flaws that are passed on to all downstream applications, and, again, fosters fairer competition.

Even among models that embrace the open-weight paradigm, keeping proprietary control over key inputs and processes creates significant, unbridgeable gaps in disclosure, which ultimately hinders scientific and ethical oversight. Kapoor et al. point out that most studies on marginal risk for common misuse vectors, such as disinformation, biosecurity and cybersecurity, are incomplete, what they propose a new framework to better evaluate the risks created by a lack of full openness.Their findings focus on two main scenarios: for Automated Vulnerability Detection in computer programs, the added risk from switching to open models is considered low, since powerful tools already exist and both attackers and defenders gain access to these new AI tools, potentially maintaining a balance between the two sides.However, in the case of Non-Consensual Intimate Imagery, the

marginal risk is demonstrably high, as open models have made it dramatically easier for non-experts to create highly realistic and targeted harmful content, overwhelming existing defences.

The study concludes with a strong call for responsibility: AI developers should be more transparent about their practices, and more independent researchers should conduct testing using updated threat models that reflect the different release practices of open-weight, open-source, and closed-source models. Policymakers should ensure that this work is properly funded and remains independent of commercial interests.

9.4.4.2 Open Weight vs Open Source

Perhaps the biggest problem to enact these recommendations is that secret training data constitute a significant barrier to research [66]. Without access, independent scientists cannot understand a model's inherent biases or attempt to replicate its training results. Reproducibility is a cornerstone of the scientific process: without access to all components, researchers cannot verify, replicate, or truly build on a model's results or on papers written by insiders at proprietary companies.

Choosing Open-Weight vs Open-Source Paradigms The difference between open-source and open-weight AI models has significant implications for development, transparency, and customisation. While often used interchangeably, they represent fundamentally different approaches to releasing AI models.

The difference between open-source and open-weight lies in which of these components are made publicly available for external inspection and which are kept as a secret of the company that curated it. True open-source models provide full access to all their components, for full transparency and reproducibility, releasing model weights, source code, training scripts and methodology, including the so-called hyperparameters, and training data (see Appendix), all under permissive licences that grant broad rights for usage, modification, and redistribution, regardless of purpose and deployment scale. Open-weight models, on the contrary, provide more limited access, mainly to the set of pre-trained weights for third-party running of the model and fine-tuning.
The technical differences between these two approaches have significant practical consequences for engineers and researchers, as can be seen from Table 9.1.

The invisible source: The data brokerage phenomenon A data broker is any entity that, in exchange for payment or other compensation, sells or otherwise makes available personal data (such as interests, demographics, behaviours, purchase history, and more) about individuals to another entity independent from the original collector. These firms act as intermediaries in the data economy, aggregating vast dossiers on millions of people and selling them for purposes like marketing, customer profiling, and analytics[1]. Their activities are legal under US law since there is no GDPR equivalent (although some states grant more rights to consumers than others, as noted above. Almost always, these transactions occur without the user's informed consent [29]; for these companies, the *consumer* is not the *customer*, but the *product*.

The process involves several stages:

	Open Source	Open Weight
Reproducibility	High. With access to the code, data, and methodology, researchers can reproduce the training process, verify claims, and gain full understanding.	Impossible. Without the training data and source code, it is impossible to recreate from scratch.
Transparency	High. Full access allows for rigorous scrutiny of the code and data that form biases and inform the model's choices.	Limited. Without access to the training data, it is difficult to thoroughly evaluate the model's potential biases and limitations.
Customisation	Deep. Developers can freely modify and adapt the model up to minuscule details.	Superficial. Customisation is primarily limited to fine-tuning the provided weights for a specific task. You can adjust the model but cannot fundamentally change it.
Control	High. Users have complete control and are not dependent on the original creator's platform or tools.	Low. The user is dependent on proprietary tools and decisions from the creators. Open Weight enables broad application development by lowering the barrier to entry for using advanced AI, but concentrates control among the original creators.

Table 9.1

Collection: Brokers acquire data from a multitude of sources. This includes scraping public records, purchasing data (e.g. transaction data) from other companies, browsing and search history via cookies, and embedding software development kits (SDKs) into third-party mobile applications.

Aggregation/Profiling: Collected data is aggregated and synthesised to build detailed profiles of individuals. These dossiers can include a wide array of deeply personal information, such as name, address, income, education, purchases, interests, political preferences, health information, legal trouble they have incurred, credit history and real-time location data [28]. What is worse, sometimes these profiles are put erroneously in the wrong category, which can have severe downstream consequences.

Monetisation: the value of data is not inherent but is derived from how it enables and improves the business decisions of the marketers and other clients who purchase it[128]. Curated datasets and profiles are exploited through two primary pathways, one direct, and the other indirect:

Indirect monetisation involves leveraging in-house collected consumer data to improve a firm's own products or services. For instance, streaming platforms like Netflix use viewing history to personalise content recommendations, thereby increasing user engagement and subscription revenue. This approach enhances the firm's own market competitiveness and improves the firm's existing revenue streams.

Direct monetisation involves the direct sale of data to third parties, which is the core business model of data brokers. The preferred places where this exchange takes place are online venues known as data marketplaces, which bring together data owners, brokers, and consumers to facilitate the commodification of data and generate synergies, in which resources from multiple brokers are combined by downstream firms to improve their accuracy and insight. [128]

The business is immensely lucrative, with the global data broker market generating hundreds of billions of U.S. dollars annually[50]. Prominent actors in this industry include Epsilon, Equifax, Acxiom, and Experian (the latter having an establishment in the EU), which maintain massive databases on hundreds of millions of consumers. It is easy to imagine how the invasive and powerful practices of the data brokerage industry pose significant risks to privacy, civil rights, and national security; the US Federal Trade Commission's (FTC) enforcement action against the the prominent location data broker X-Mode Social, Inc., provides a concrete illustration of the risks: before it was prohibited from engaging in the sale or sharing of sensitive location data, their SDK was embedded in over 300 apps, including gaming, health trackers, and religious apps, allowing it to gather over 10 billion location data points daily, with 70% of the data having a location accuracy within 20 meters. An investigation by the American Federal Trade Commission revealed that despite users opting out of personalised advertising, X-Mode continued to collect and sell their data, while they completely omitted to inform the public about the potential use of their data for national security purposes by Government contractors. It was also ascertained that the company created audience segments based on sensitive

attributes, such as visits to medical facilities or religious sites [129]. Meanwhile, another broker, *Near Intelligence*, reportedly sold location data from visits to Planned Parenthood clinics to anti-abortion activists [17].

The connection between the data brokerage industry and the sourcing of training data for AI models is direct and symbiotic, and is one of the most critical—and problematic—aspects of the modern AI ecosystem. Data brokers are a primary, though often unacknowledged, part of the supply chain for the vast datasets needed to build and train powerful AI systems, especially foundation models. The power of LLMs is inherently tied to the scale of their training data[105]. Massive datasets containing billions of words are essential for effective pretraining, exposing the model to the vast diversity of language in real-world contexts. As performance improves with more input data, there is a constant and enormous demand for massive, diverse datasets. Thanks to machine learning, this data does not need to be pre-processed, labeled, or curated in advance. Data brokers specialise in this activity, systematically collecting and curating vast repositories of data from disparate sources and providing exactly the kind of large-scale, granular data that is ideal for training AI models—especially those designed to predict human behaviour, preferences, and needs.

However, the AI development pipeline often obscures the origins of its training data. From the final product alone, it is nearly impossible to trace a specific output back to a particular piece of training data. Once the source is identified, it is even more difficult to remove its influence from the model's weights. This process is known as data laundering: data collected unethically or without informed consent by a broker is absorbed into a model, which is then presented as a new, «clean» technological product with no practical way to trace it back to the original, potentially illegitimate, source.

Just as cause and effect are blended, so is the transference of risk from broker to model. Training data may be inaccurate or reflect a more negative and problematic reality than truly exists. These risks are not only inherited by the model, but also amplified and better concealed.

Anyone familiar with the principles of the GDPR can see that the data brokerage model fundamentally conflicts with data minimisation, consent, and the right to erasure or correction of wrong information, as it becomes nearly impossible to honour once their data is integrated into a trained model. Even adding the possibility to subjects to intervene in their specific cases does little to address the larger issue: data broker datasets often crudely and uncheckedly reflect existing societal inequalities by over-representing some demographics and under-representing or misrepresenting others.This is especially problematic when proxies are used to make decisions in the absence of more relevant data. The adage «Garbage In, Garbage Out» captures the problem: inaccurate datasets, which may blend the identities of people with the same names or contain outright errors, lead models to learn false correlations and factual inaccuracies.This significantly contributes to the phenomenon of hallucination, where models produce confident-sounding but nonsensical outputs.And because many data brokers source from basically the same internet spaces, even different models can fall prey to the same blind spots and poison countless downstream applications. The European Commission has highlighted that when an AI model is trained on such data, it learns and amplifies these biases, resulting in discriminatory outcomes in areas like hiring, loan approvals, and

criminal justice.

In response to these challenges, researchers have called for greater transparency throughout the data lifecycle. Proposals include requiring datasheets and model cards documenting the provenance, composition, and known limitations of training data and the resulting models, as well as creating data *hubs* to better support data management, curation, documentation, and quality assessment across the foundation model lifecycle.

Chapter 10

Case Studies

This chapter presents a series of representative cases that illustrate the various pathways toward achieving compliance with the AI Act when AI systems are deployed in real-world contexts.

10.1 Closed-Source Algorithms and determining compliance

A recurring challenge in the legal tech discipline is assessing the viability of potential high-risk products from grassroots startups. In the eyes of an engineer accustomed to clear-cut specifics and rigid protocols, the initial impact in these cases could be one of confusion, due to a lack of detail and uncertainty about the product's inner workings. The AI Act primarily examines **purposes** and **declared uses**, but the actual, implicit goals that a specific model will pursue are hard to know without full access to every nook and cranny of the model used. If the model is a closed source one, then what is known stops at what can be loosely inferred from the prompts going towards to external, proprietary APIs.

The extreme success of products like ChatGPT, combined with the great support that EU and national grants are providing to innovators, has encouraged a significant proliferation of enterprises that offer B2C, or, more often, B2B services in Europe. In the eyes of the Act, these startups are *deployers* of a powerful, general-purpose AI model developed by a tech giant, but they are *providers* of the service built on top of it, which could fit the definition of high-risk for the AI Act. They would therefore bear the *provider*'s responsibility for their product's output, **while lacking full access and control over the underlying model** (essential to fulfilling the Act's requests meaningfully): they cannot, for instance, produce documentation on the training data. This **information asymmetry** places them in a precarious position, as they build their entire business around a *black box*.

A consultant hired by them to determine compliance with the AIA will need information from the startup regarding the product's robustness, accuracy, and cybersecurity, among all the full requirements to draft a detailed risk management system, as mandated by *Article 9*. In such a scenario, the startup itself will often only have access to the **terms of service** of their provider, with no possibility to request detailed auditing.

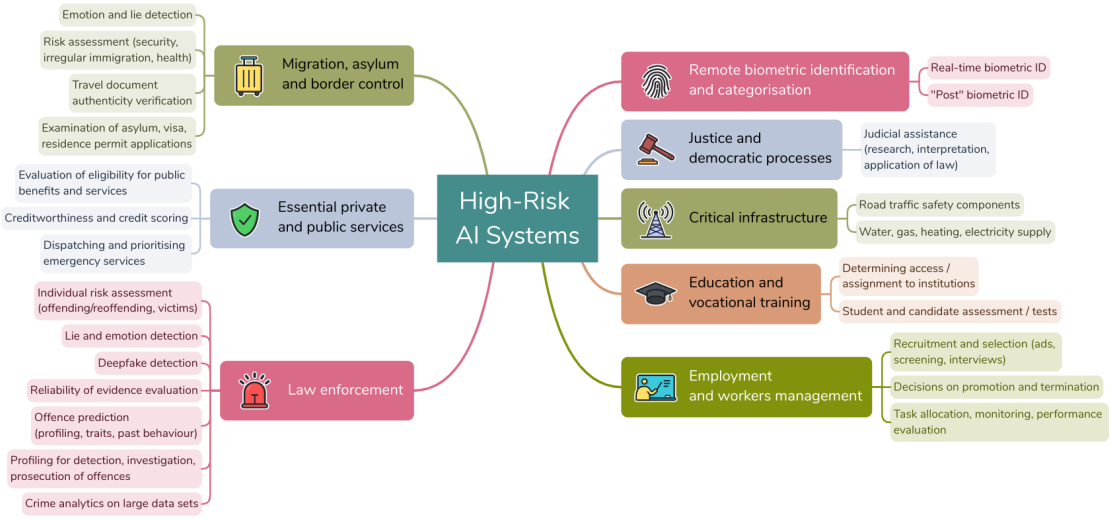


Figure 10.1: Summary of High-risk AI systems pursuant to *Article 6(2)*, as listed in *Annex III* of the AI Act.

They only have an «*illusion of control*» over their product: their only intervention can be **prompt engineering** and **post-processing filters**.

This ex-post, surface-level approach stands in stark contrast to the Act's requirement for ***security-by-design***, leaving a significant portion of the actual risk undocumented and the second-to-last-mile user to bear the consequences of any potential issues.

10.1.1 The need for documentation

Creating and/or using a general-purpose model, such as any of the top products present in the market at the moment[1] attracts consistent compliance responsibilities.The AI Act requires, in *Article 53*, that **providers of general-purpose AI** models **publish** the following information[2]:

> *1. A general description of the general-purpose AI model including:*
> *(a) the tasks that the model is intended to perform and the type and nature of AI systems into which it can be integrated;*
> *(b) the acceptable use policies applicable;*
> *(c) the date of release and methods of distribution;*
> *(d) how the model interacts, or can be used to interact, with hardware or software that is not part of the model itself, where applicable;*
> *(e) the versions of relevant software related to the use of the general-purpose AI model, where applicable;*
> *(f) the architecture and number of parameters;*
> *(g) the modality (e.g. text, image) and format of inputs and outputs;*
> *(h) the licence for the model.*

[1] Including OpenAI's *GPT*, Google's *Gemini*, and Anthropic's *Claude*.
[2] Annex XII.

2. A description of the elements of the model and of the process for its development, including:

(a) the technical means (e.g. instructions for use, infrastructure, tools) required for the general-purpose AI model to be integrated into AI systems;

(b) the modality (e.g. text, image, etc.) and format of the inputs and outputs and their maximum size (e.g. context window length, etc.);

(c) information on the data used for training, testing and validation, where applicable, including the type and provenance of data and curation methodologies.

Article 55 then adds the to the list of obligations the following ones for those that embody a **systemic risk**:

(a) perform model evaluation in accordance with standardised protocols and tools reflecting the state of the art, including conducting and documenting adversarial testing of the model with a view to identifying and mitigating systemic risks;

(b) assess and mitigate possible systemic risks at Union level, including their sources, that may stem from the development, the placing on the market, or the use of general-purpose AI models with systemic risk;

(c) keep track of, document, and report, without undue delay, to the AI Office and, as appropriate, to national competent authorities, relevant information about serious incidents and possible corrective measures to address them;

(d) ensure an adequate level of cybersecurity protection for the general-purpose AI model with systemic risk and the physical infrastructure of the model.

OpenAI's models are an example of models which are classified as posing **systemic risks** in relation to these requirements. A comprehensive peer review [48] of the GPT-4 technical report noted a stark «*lack of clarity in training processes*» as a significant limitation to determine compliance, raising concerns about hidden, encoded «*biases and interests*». More than 100 leading AI researchers signed an open letter calling for companies like OpenAI to allow access to their systems for safety testing, but OpenAI cites the need to protect the model from both competitors and potential abusers [89] as the reason they have been closed off until now.

10.2 Case Evaluation: High-Risk AI in Recruitment

The compliance evaluation of an AI-based curriculum vitae (CV) assessment tool ultimately hinges on *Article 5* (*Prohibited AI Practices*) and *Annex III* of the Artificial Intelligence Act, which prescribes the criteria for **system prohibition or high-risk classification**. Presume that the client presents the tool as utilising a widely known open-weight Large Language Model, with **no additional fine-tuning**, to systematically process a significant volume of CVs submitted by prospective employees, extracting

explicit data points such as *skills, work experience, education, and contact details*. Additionally, it infers **implicit** skills and assigns proficiency scores ranging from 1 to 5. The evaluation further considers **previous employer data** extracted from an external database to refine candidate profiles. Upon request, the software identifies candidates whose computed **scores** align with the requirements of the job description.

An initial, higher-level analysis of this system reveals a definite distance from most of the prohibited practices as per Article 5, except for the case of paragraph c:

> *[The following AI practices shall be prohibited:]*
>
> *[...]*
>
> *(c) the placing on the market, the putting into service or the use of AI systems for the* ***evaluation or classification of natural persons*** *or groups of persons* ***over a certain period of time*** *based on their social behaviour or* ***known, inferred or predicted personal or personality characteristics****, with the social* ***score*** *leading to either or both of the following:*
>
> *(i) detrimental or unfavourable treatment of certain natural persons or groups of persons* ***in social contexts that are unrelated to the contexts in which the data was originally generated or collected****;*
>
> *(ii) detrimental or unfavourable treatment of certain natural persons or groups of persons that is* ***unjustified or disproportionate to their social behaviour or its gravity****;*
>
> (emphasis added)

It is clear from the description that the system under examination is designed to evaluate individuals based on **known, inferred, and predicted personal characteristics**, with the consequential assignment of **scores** that, in certain instances, may lead to unfavourable outcomes (being excluded from the job). However, the requirement that such an evaluation be conducted over an extended period appears inconsistent with the software's **one-time utilisation** to select a particular candidate. Furthermore, the use of **static data**—namely, a definitive list of past experiences and current personal details—further supports this interpretation. Any unfavourable outcome arises exclusively from the analysis of data provided in **relevant contexts**, such as CV submissions or public registries of former employers, which bear a direct relation to the evaluation context. Notably, the individual subject voluntarily prepares and submits their CV for this assessment.

The biggest point in the case of a favourable opinion of this application is the fact that the *unfavourable* treatment is not supposed to be realised in social contexts that are *unrelated* to the ones in which the data was originally collected, and an eventual negative result will concern only the related job posting.

Nevertheless, several further considerations warrant attention:

- Although the tool may not fall under the AI Act's prohibited practices (*Article 5*), it is likely to be classified as a *high-risk* AI system under *Article 6*, due to its involvement in **employment-related functions** such as **recruitment and**

candidate evaluation, which are explicitly referenced in *Annex III*. While *Article 6(3)* outlines **exceptions**, particularly for tools that serve as aids to predominantly manual tasks, these exceptions do not apply to systems involving profiling. As of this writing, the European Commission has not yet issued further official guidance or clarification on these rules.

- **Information *enrichment*** derived from scraped data can be highly **imprecise** and **outdated**, particularly when limited information is available online or when names are ambiguous. Although the European Union has restricted data brokerage under the GDPR, the tool's deployment in countries such as the United States, where data reselling is less regulated, would likely lead to reliance on poorly curated data sources. This issue has been extensively documented [80, 31, 121], and its resolution typically requires manual correction, which the tool is designed to avoid.

- If the software were used by intermediaries who act as middlemen between prospective employees and employers, and who durably maintain candidate profiles in their databases across multiple jobs and opportunities, the chance that the profiles would be kept in-between working relationships is astronomically high. Indeed, those who look for employment but are frequent job hoppers or only find limited-time opportunities will **remain in the candidate pool for many searches and comparisons**. Although the risk would be mitigated by the coherent context across uses, which would also be consistent with the reason the subjects had surrendered their data, the potential for sneakily **profiling across multiple scenarios** is still there.

- Scraping data without time restrictions may result in *de facto* historical profiles as if candidates had been monitored before their applications.

- The type of inferences generated by the model depends on its training data and configuration. Although the Llama model's open weight architecture provides some insight into its decision-making process, transparency into its training data remains incomplete. The lack of fine-tuning, particularly in the context of an American-developed model, raises concerns about embedded biases identified in recent literature. A more prudent approach would be to use a fine-tuned, restricted model that **relies exclusively on vetted, impartial sources, such as the Companies Register of the Italian Chamber of Commerce**.

In summary, although an AI-based CV assessment tool does not fall within the prohibited practices described in Article 5 of the AI Act, it constitutes a high-risk system due to its use in recruitment and candidate profiling. Moreover, the tool's reliance on inferred skills, outsourced data enrichment, and the possibility of outdated or biased information present substantial concerns, especially with respect to transparency, fairness, and data protection. As regulatory frameworks evolve, it is imperative that providers mitigate these risks through strengthened oversight, increased model transparency, and more robust data sourcing.

10.3 Deepfakes for employee training

For a company, producing professional-looking videos is rarely an inexpensive task. At a minimum, it requires high-quality equipment, a dedicated soundproof recording space, and time to draft scripts, record them (often with multiple takes if the speaker makes a mistake), and an editor.

With the advent of AI-generated deepfakes, many of these challenges can potentially be solved. Major companies in this field already promise substantial time and cost savings, delivering «*studio-quality*» videos featuring avatars that can speak over a hundred languages at the same time. Avatars never need rest, can be customised to fit the cultural norms of their audience, and are able to confidently discuss any topic. They are well-suited for advertisements, employee onboarding guides, security training, and even tutorial videos.

Despite the wide variety of available avatars, an employer might prefer even greater personalisation and choose to feature a familiar face—such as a chief of staff or HR coordinator. Before doing this, however, it is imperative to reflect on the possibility that the practice of using a regular employee's likeness for AI training is risky in terms of their fundamental rights. A plan must be in place to avoid prohibited or high-risk situations for employees before the project can even begin.

Risk classification Creating a realistic 3D model requires collecting extensive information about the subject's appearance. Since the algorithm must simulate natural behaviour, it may also need to recognise and reproduce emotions as required by the script. Despite these considerations, this application of AI would probably be classified as presenting **minimal risk** , at least as far as the AI Act is concerned. the deepfake algorithm's purpose is simply to collect a set of photos and/or videos to learn how to replicate them. This process cannot really be considered **surveillance**, as it is not **continuous in time**, nor is the content used to make decisions or judgements about the individual. However, this application is still an AI system designed to interact with humans and generate synthetic content. As a result, it is subject to **transparency requirements** outlined in Article 50, which ensure that people are informed when they are interacting with AI or viewing artificially generated content. This helps prevent confusion and maintains ongoing involvement of the person whose likeness is being used, ensuring they are always kept informed about how their image is being utilised (as per Recital 165).

The remaining points of interest and the exceptions listed in Annex III do not appear to be relevant in this case.

When going outside the purview of the AI Act, it is clear that this case still raises **a data protection concern**, since the employee's likeness will be stored in detail and potentially distributed to a wide audience. If this data were leaked to unauthorised third parties, they could use it to train their own models, creating lifelike reproductions of real individuals who might say or do things that could harm the original person's reputation. Similarly, this likeness could be misused to deceive others, for example, in sophisticated spear-phishing campaigns.

The danger of deepfakes Several high-profile criminal cases have already involved *deepfakes* being used for phishing and other malicious purposes, resulting in severe financial, reputational, and psychological consequences. This can be seen also as a human rights issue, as family emergency scams [55, 4] feigning the arrest or injury of a loved one or the creation of non-consensual pornography can have scarring effects on an individual, but also one that concerning the financial and reputational well-being of a company. There is no scarcity of cases in which corporate phishing and large-scale financial fraud have been successfully carried out using sophisticated *avatars* impersonating high-level executives to convince and authorise fraudulent transactions. For example, in 2024, a finance employee at Arup's (a multinational design and engineering firm) Hong Kong branch was deceived into transferring 25,6 million dollars after a «company executive» convinced an employee such request was genuine [78]. Similarly, in 2019, the CEO of the UK branch of an energy firm was tricked by an AI-generated voice mimicking his parent company's boss, leading to a 243000 dollars transfer. The attacker used AI to clone the voice of the German CEO, down to his subtle accent and vocal patterns [77]. In yet another case, scammers used a fake WhatsApp account and a cloned CEO voice to target the global firm *WPP*, but the attempt was foiled by suspicious employees [[77]].

Having assessed the dangers of a deepfake model in the wrong hands, it is clear how raw training data should be kept only for as long as absolutely necessary, and how important it is to secure all devices, accounts, and networks that can access the trained model using state-of-the-art protection methods. In some cases, conducting a Data Protection Impact Assessment (DPIA) may also be appropriate.

10.4 AI Invoice Ads Automation

Artificial Intelligence primarily offers efficiency and seamless data processing. For instance, automated pipelines can replace the manual input of supplier invoices into databases. Such applications are generally considered **outside the critical risk framework defined by the AI Act**, even when highly autonomous and minimally supervised, because they represent a low risk to humans (an error would typically only impact a company's finances). If we consider a case where the main problem the AI agent is supposed to solve is the scraping of individually received PDF documents that are, at that moment, visually scanned by an human employee, and then transcribed, it is obvious that this process is set to reduce enormous overhead and a great loss in human time, while not feeling the loss of a *human* input. On the contrary, considering the state of the art for AI-augmented OCR technologies, human intervention would be a significant slowdown. The implementation of AI tools for *menial tasks* like transcribing invoice data into databases has been so widely studied that nowadays it has overwhelmingly positive impact on productivity, greatly offsetting any kind on (minimal) risk derived from the use of AI.

10.4.1 The benefit of introducing AI in accounting

Artificial Intelligence Optical Character Recognition (AI-OCR) technology is already revolutionising *Accounting Information Systems* (AIS), particularly within *Vendor Invoice Management Systems*. AI-OCR not only digitises text but also enables context understanding, manages unstructured data, learns from past iterations, and continuously improves its accuracy over time. Future directions anticipate further modernisation, including integrating AI-OCR with technologies such as Natural Language Processing for advanced financial analysis.

Modular and configurable AI-based platforms allow[116] rapid deployment across a range of clients with minimal effort. These platforms let businesses set **specific rules, thresholds, and validation steps** (such as line-item matching, and tax code determination) to be set via intuitive interfaces. This flexibility makes them suitable for diverse industries and compliance needs, addressing a key limitation of earlier rule-based solutions, which could only manage **repetitive**, **structured** data and struggled with unstructured information or industry-specific **exceptions**. The OCR technologies of yesterday were markedly imperfect due to their rigidity in parsing complex, variable layouts and had next to none ability to adapt to fringe cases; now, they can handle catalogue mismatches and use semantic similarity, lexical normalisation (e.g. transforming abbreviations in the full word), and fuzzy matching (to perform similarity analysis and find near-matches or transpositions, e.g., INV-NO-123 vs. INV123, or 1908 vs. 1980, a clear input mistake) to reconcile vendor, product, and line-item apparent mismatches[116], enabling robust processing. And that is not all: by comparing vendor names, invoice numbers, amounts, and dates[59] with historical records, and using techniques like fuzzy matching and string similarity metrics, these systems can mitigate financial leakage from duplicates, addressing not only human error but also **intentional fraud**.

Finally, as AI takes on tasks like extraction, parsing, organisation, and automated verification (e.g. matching invoices to purchase orders), it can also be directed to create a complete **audit trail**. This capability enables **proactive error detection**, improved **compliance** with rules (both internal and external), and greater **transparency** in financial processes. Firms serviced by AI-using accountants demonstrated increased ledger granularity, evidenced by a 12% increase in unique general ledger accounts [23].

Measurable results Insight from the IT world[71] shows without doubt that the integration of AI-OCR into AIS has been revolutionary, leading to significant improvements across various accounting functions in efficiency, speed, accuracy, and cost-effectiveness (for example, a typical implementation using Google Cloud's Document AI service, costing roughly $150 per month, can deliver a projected return on investment of up to 40 times its cost[99].). Case studies focused on the old way of doing things report an expense of about 25% of total working hours spent just processing invoices[99], a task that is fundamentally straightforward but fraught with opportunities for error. In contrast, AI solutions want to minimise the need for human intervention, which should only be required as a last resort.

End-to-end automated invoice processing systems, developed using AI modules, have been demonstrated to compress end-to-end cycle time (from receiving to posting invoices) by more than 90%[116], with one system successfully processing approximately

80000 invoices for two clients, achieving fully automated processing for 76% of them [59], in the face of very little work to configure it at the beginning.

Instead of replacing human expertise entirely, AI allows professionals to **shift away from low-value, repetitive tasks toward higher-value activities**, **augmenting, not replacing, accountant expertise**: recent studies on AI adoption in accounting firms document accountants reallocating approximately 8.5% of their time from routine data entry to tasks such as business communication and quality assurance [23]. The resulting increased capacity allows teams to focus resources on **strategic activities** and handle growing invoice volumes without increasing headcount. The more experienced accountants demonstrated **strategic complementarity by selectively overriding AI suggestions**, but, at the same time, a framed field experiment[23] highlighted **the risk of passive use**, showing that people sometimes over-rely on inaccurate AI-suggested classifications, a risky behaviour considering that independent benchmarks[9] and comparisons of state-of-the-art models (like *Yolov8x*) report a maximum accuracy of 91%, with results as low as 53%. Nevertheless, natural language processing and intelligent OCR have been steadily improving reliability compared to error-prone human input, , which in turn will help save 500000 to 2.5 million US Dollars annually[116] .

All in all, the application of AI on automating invoicing operations is a practice that could become a standard example for an application that is safe for the humans involved and has been observed to improve significantly productivity and competitiveness of a firm (although **effective oversight by human expertise** remains crucial to mitigate the risks of propagating errors), a prime example of the **types of applications that the AI Act encourages and is here to foster**. Unfortunately, this conclusion cannot be extended haphazardly to the whole financial sector, due to the potentially sensitive nature of the data involved.

Chapter 11

Towards models for the common good

AI has been demonstrated to be able to support a fairer society. Data scientist **Cathy O'Neil** and product designer **Richard Pope** each explain in their respective books[100, 80] how AI can serve ethical goals and the public good, rather than focusing solely on profit. Both argue that **the impact of AI depends on the values and objectives embedded within it**. When priorities abandon pure efficiency, AI models can become tools for providing meaningful support and justice.

11.1 Housing assistance

The clearest and most detailed example comes from O'Neil's work: a model to reduce homelessness. She describes her role as «*helping to develop a recidivism model*» similar to those used in the American judicial system to predict which individuals are most likely to reoffend. This time, however, the goal is to understand the «*forces that pushed people back to shelters and those that helped them achieve stable housing*», in order to help address homelessness. Ultimately, the model identified a key factor: people who did not return to homeless shelters after leaving were those who had received housing assistance through the *Section 8* program. This insight was quickly censored by public officials, who at that time were actively working to move families off this effective program as part of their policies. This demonstrates how **even a well-designed, beneficial model can be rendered ineffective if policy priorities are driven by perceived cost savings rather than actual impact**.

11.2 Preventing child abuse

Cathy O'Neil also presents another example of a model designed with the common good in mind: one that identifies households where children may be at higher risk of abuse (such as families with members with a history of drug use or domestic violence). On the surface, the model functions like many problematic AI systems, using historical data to **find statistical correlations that predict future outcomes**. However, the crucial difference is that **the goal is not to punish parents who fit the profile**

identified by the model, but rather to provide them with additional resources and support—such as after-school programs and counselling—before any harm occurs.

O'Neil's experience shows again that a model's objective is what transforms it from potentially harmful into genuinely beneficial: the same system, if programmed differently, could have been used to target vulnerable families for predatory marketing or to remove their children preemptively. This is a clear example of why every model requires careful oversight, explicit and auditable decisions about its deployment, and ongoing transparency regarding its intended outcomes.

11.3 A cornerstone example for AI in public administration

In *Platform Land*, Richard Pope critiques the current state of **digital public services** and proposes a new vision in which **the government itself acts as a platform**. He argues that simply moving paper forms online has failed to address the fundamental problems of bureaucracy. Instead, this approach has fragmented processes and obscured how things work, placing burdens on citizens who struggle to **discover available services, provide required forms and evidence**, and cope with the anxiety and stress caused by uncertainty. Digital services have simply swapped paper-based hassles for digital ones, overlooking the persistent gaps and overlaps between various services where much of the complexity remains. As a result, the government becomes harder to understand and less accountable for how it handles each individual's case and needs.

These issues are not just technical glitches, but often stem from using technology to enforce explicit yet flawed **policy choices**. The UK's *Universal Credit* system illustrates this: assumptions built into the policies—such as people being paid monthly, couples sharing finances, or job-searching tasks reliably leading to employment—created a system that forced users into debt and confusion. Pope argues that most government digital transformation efforts have just focused on maintaining « *the status quo, delivered more cheaply* » rather than reducing the bureaucratic burden on citizens.

The proposed alternative is a fundamental **rethinking of how services are built, managed, and delivered**, using **shared digital infrastructure** to be more **proactive**, humane, and accountable. The author calls this approach « *Government as a Platform* »," a new relationship between the citizen and the State offering **proactive support**. This system leverages AI algorithms similar to those used by e-commerce and streaming platforms to suggest products and movies to watch, but this time for a different kind of audience: the algorithm identifies a citizen's needs (such as applying for social housing or enrolling a child to school) and automatically provides the right forms and information whenever relevant, using all available knowledge about the individual. The result is **services that adapt in real time to life changes**, automating renewals, surfacing recommendations, and pre-filling applications. For example, the system could detect when a child is approaching school age and automatically prompt the parent to start the enrolment process.

The recommended approach is to treat foundational data (like addresses, business

records, or land ownership) as shared infrastructure, accessed through APIs and served from a single authoritative source whenever needed. From an architectural perspective, this system is made up of **modules that work closely together** but remain **independent**, ensuring both scalability and adherence to the «once-only» principle: data provided to one government service should be reusable across others (**with user consent**), eliminating the need for citizens to provide the same information repeatedly. *Data duplication* is seen as one of the worst anti-patterns for complex platforms, as it inevitably leads to the impossibility of keeping records **updated and correct** across every corner of the government (a cornerstone of European data protection law).
Unfortunately, while Pope's vision promises radical simplification for users, it also highlights a paradox: **simplification can create new types of complexity, especially through the introduction of powerful algorithms that allow many things to happen simultaneously and in real time**. The challenge is to **keep users** oriented and **in control** without overwhelming them, establishing loops that consist on «bursts» of automation followed by tasks performed by a human, and keeping clear, auditable logs that explain the past, present and future of each piece of information belonging to the citizen (when it is accessed, and why, complete with a list of steps generated and prompted by automated checks and interventions).
Government as a Platform is both a technical and political project. Pope anticipates that it will **shift power dynamics** and introduce **new challenges**. Because of deep personalisation, no two users will have exactly the same experience, requiring **new methods for auditing** whether the system is functioning correctly when there is **no single standard response**. This complexity is inherent in AI and, more broadly, in modern software development, which evolves continuously and often depends on outdated or incomplete documentation, which hinders effective oversight. Developers may use ongoing changes and personalisation as reasons to withhold details, making systems less transparent and harder to scrutinise. If a journalist or politician asks how the system works, officials can claim there is no single answer.

To address this, the author proposes «*automated transparency*»: systematically publishing information about **system operations, code, version histories, and data usage policies**, through four main steps:

- Making the **inner logic and rules** that govern the system inspectable and public;
- Automatically publishing version histories, change logs, and design archives to keep a record of how and when the system has **evolved**.
- Providing users with a full disclaimer on how and why their individual pieces of data are used. A journal-type pattern is implemented whenever it is appropriate to show a user their complete interaction history and the path their data has taken through services.
  A key feature of the harmful WMDs mentioned above is that the data they use is secret, and their decisions are beyond dispute or appeal; in contrast, *automated transparency* mandates that data usage be made visible, for both individuals and public bodies.
- **Explaining** decisions, a.k.a. *Transparency at the point of use*: digital services provide explanations in context, as the decisions are made, allowing users to see

the calculation behind them, and the data that fuelled them, so that anybody can pinpoint eventual errors and give timely and precise feedback.

Richard Pope's vision for a fairer digital public administration acknowledges major challenges and highlights the need for renewed commitment to **transparency, accountability, and the responsible use of power**. Still, this approach offers the possibility of a state that is not only more efficient, but also more understandable, just, and democratic. By embedding **transparency**, **intervenability**, and **proactivity** at the core of digital government, these values become fundamental to its operation and can help fulfil the promise of truly equitable governance.

Bibliography

[1] Laura Abrardi, Carlo Cambini and Flavio Pino. *Data Brokers Competition, Synergic Datasets, and Endogenous Information Value*. Available at SSRN: https://ssrn.com/abstract=4901441. July 2024. URL: https://ssrn.com/abstract=4901441.

[2] Technology Action Group on Erosion and Concentration. *BehindSugarandSpiceandEverythingNic TheEnvironmentalImpactsofDigitalization. Communiqué119*. Accessed: 2025-11-09. 2024. URL: https://www.etcgroup.org/sites/www.etcgroup.org/files/files/sugar_and_spice_word_final.pdf.

[3] *AI Risks Check-list for AI Risks Management*. AI Standards Hub. URL: https://aistandardshub.org/ai-standards/ai-risks-check-list-for-ai-risks-management/ (visited on 18/11/2025).

[4] Mahrus Ali et al. "Deepfakes and Victimology: Exploring the Impact of Digital Manipulation on Victims". In: *Substantive Justice International Journal of Law* 8.1 (May 2025). ISSN: 2599-0462. DOI: 10.56087/substantivejustice.v8i1.306. URL: http://dx.doi.org/10.56087/substantivejustice.v8i1.306.

[5] Marco Almada and Anca Radu. "The Brussels Side-Effect: How the AI Act Can Reduce the Global Reach of EU Policy". In: *German Law Journal* 25.4 (Feb. 2024), pp. 646–663. ISSN: 2071-8322. DOI: 10.1017/glj.2023.108. URL: http://dx.doi.org/10.1017/glj.2023.108.

[6] *AP and OpenAI Agree to Share News Content and Technology*. The Associated Press. 13th July 2023. URL: https://apnews.com/article/openai-chatgpt-associated-press-ap-f86f84c5bcc2f3b98074b38521f5f75a (visited on 18/11/2025).

[7] Judith Arnal. "AI at Risk in the EU: It's Not Regulation, It's Implementation". In: *European Journal of Risk Regulation* (Mar. 2025), pp. 1–10. ISSN: 2190-8249. DOI: 10.1017/err.2025.19. URL: http://dx.doi.org/10.1017/err.2025.19.

[8] John Bagby and Nizan Packin. "RegTech and Predictive Lawmaking: Closing the RegLag Between Prospective Regulated Activity and Regulation". In: *Michigan Business and Entrepreneurial Law Review* 10.2 (2021), p. 127. ISSN: 2375-7523. DOI: 10.36639/mbelr.10.2.regtech. URL: http://dx.doi.org/10.36639/mbelr.10.2.regtech.

[9] Merxhan Bajrami et al. "Deep Dive into Invoice Intelligence: A Benchmark Study of Leading Models for Automated Invoice Data Extraction". In: *Proceedings of Ninth International Congress on Information and Communication Technology*. Ed. by Xin-She Yang et al. Singapore: Springer Nature Singapore, 2024, pp. 177–191. ISBN: 978-981-97-3289-0.

[10] Adrien Berthelot et al. "Understanding the environmental impact of generative AI services". In: *Communications of the ACM* Special Issue on Sustainability and Computing 68.7 (2025), pp. 46–53. DOI: 10.1145/3725984. URL: https://hal.science/hal-04920612.

[11] Jeremy Bertomeu et al. "Capital Market Consequences of Generative AI: Early Evidence from the Ban of ChatGPT in Italy". In: *SSRN Electronic Journal* (2023). ISSN: 1556-5068. DOI: 10.2139/ssrn.4452670. URL: http://dx.doi.org/10.2139/ssrn.4452670.

[12] JULIA BLACK and ROBERT BALDWIN. "Really Responsive Risk-Based Regulation: REALLY RESPONSIVE RISK". In: *Law and Policy* 32.2 (Mar. 2010), pp. 181–213. ISSN: 0265-8240. DOI: 10.1111/j.1467-9930.2010.00318.x. URL: http://dx.doi.org/10.1111/j.1467-9930.2010.00318.x.

[13] C. Boine and D. Rolnick. "Why the AI Act Fails to Understand Generative AI". en. In: *SSRN* (2023). URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4644701.

[14] Rishi Bommasani et al. *The 2024 Foundation Model Transparency Index*. 2024. DOI: 10.48550/ARXIV.2407.12929. URL: https://arxiv.org/abs/2407.12929.

[15] Katherine Bourzac. *GenerativeAIHasaMassiveE-WasteProblem*. Accessed: 2025-11-09. 2024. URL: https://spectrum.ieee.org/e-waste.

[16] Anu Bradford. "The False Choice Between Digital Regulation and Innovation". In: *SSRN Electronic Journal* (2024). ISSN: 1556-5068. DOI: 10.2139/ssrn.4753107. URL: http://dx.doi.org/10.2139/ssrn.4753107.

[17] *Cellphone location data used to target abortion misinformation to visitors at 600 reproductive health clinics in 48 states*. Accessed: 2025-10-19. ACLU Massachusetts. Feb. 2024. URL: https://www.aclum.org/en/press-releases/cellphone-location-data-used-target-abortion-misinformation-visitors.

[18] European Commission. Joint Research Centre. *Analysis of the preliminary AI standardisation work plan in support of the AI Act*. LU: Publications Office, 2023. DOI: 10.2760/5847. URL: https://data.europa.eu/doi/10.2760/5847.

[19] Aaron Chatterji et al. *How People Use ChatGPT*. w34255. Cambridge, MA: National Bureau of Economic Research, Sept. 2025, w34255. DOI: 10.3386/w34255. URL: http://www.nber.org/papers/w34255.pdf (visited on 24/09/2025).

[20] Foo Yun Chee. *Apple to Delay Launch of AI-Powered Features in Europe, Blames EU Tech Rules*. Reuters. 21st June 2024. URL: https://www.reuters.com/technology/artificial-intelligence/apple-delay-launch-ai-powered-features-europe-blames-eu-tech-rules-2024-06-21/ (visited on 18/11/2025).

[21] Hailin Chen et al. *ChatGPT's One-year Anniversary: Are Open-Source Large Language Models Catching up?* 2023. eprint: arXiv:2311.16989.

[22] P.G. Chiara. "Italian DPA Fines OpenAI for GDPR Non-Compliance: The Last Episode of the Garante – OpenAI Saga?" In: *European Data Protection Law Review* 11.1 (2025), pp. 102–108. ISSN: 2364-284X. DOI: 10.21552/edpl/2025/1/17. URL: http://dx.doi.org/10.21552/edpl/2025/1/17.

[23] Changyong Choi and Jihye Yoon. "AI policy in action: the Chinese experience in global perspective". In: *Journal of Policy Studies* 40.2 (June 2025), pp. 1–23. ISSN: 2800-0714. DOI: 10.52372/jps.e685. URL: http://dx.doi.org/10.52372/jps.e685.

[24] Christian Clemm et al. *Towards Green AI: Current status and future research*. 2024. arXiv: 2407.10237 [cs.CY]. URL: https://arxiv.org/abs/2407.10237.

[25] Mike Conover et al. *Free Dolly: Introducing the World's First Truly Open Instruction-Tuned LLM*. Blog post. Apr. 2023. URL: https://www.databricks.com/blog/2023/04/12/dolly-first-open-commercially-viable-instruction-tuned-llm.

[26] Data Protection Commission. *Data Protection Commission launches inquiry into Google AI model*. 12th Sept. 2024. URL: https://www.dataprotection.ie/en/news-media/press-releases/data-protection-commission-launches-inquiry-google-ai-model (visited on 23/10/2025).

[27] Pierre Dewitte. "AI Meets the GDPR: Navigating the Impact of Data Protection on AI Systems". In: *The Cambridge Handbook of the Law, Ethics and Policy of Artificial Intelligence*. Ed. by Nathalie A.Editor Smuha. Cambridge Law Handbooks. Cambridge University Press, 2025, pp. 133–157.

[28] Jasdev Dhaliwal. *What Is a Data Broker?* Accessed: 2025-10-19. McAfee Blog. Sept. 2024. URL: https://www.mcafee.com/blogs/tips-tricks/what-is-a-data-broker/.

[29] Denise DiPersio. "Selling Personal Information: Data Brokers and the Limits of US Regulation". In: *Proceedings of the Workshop on Legal and Ethical Issues in Human Language Technologies @ LREC-COLING 2024*. Ed. by Ingo Siegert and Khalid Choukri. Torino, Italia: ELRA and ICCL, May 2024, pp. 39–46. URL: https://aclanthology.org/2024.legal-1.7/.

[30] Martin Ebers. "Truly Risk-based Regulation of Artificial Intelligence How to Implement the EU's AI Act". In: *European Journal of Risk Regulation* 16.2 (Nov. 2024), pp. 684–703. ISSN: 2190-8249. DOI: 10.1017/err.2024.78. URL: http://dx.doi.org/10.1017/err.2024.78.

[31] Electronic Privacy Information Center (EPIC). *Data Brokers*. https://epic.org/issues/consumer-privacy/data-brokers/. Accessed: 2025-10-16. n.d.

[32] Warren J. von Eschenbach. "Transparency and the Black Box Problem: Why We Do Not Trust AI". In: *Philosophy and Technology* 34.4 (Sept. 2021), pp. 1607–1622. ISSN: 2210-5441. DOI: 10.1007/s13347-021-00477-0. URL: http://dx.doi.org/10.1007/s13347-021-00477-0.

[33] EU Needs AI. *Ensuring AI innovation in Europe: Open letter to EU policymakers*. URL: https://euneedsai.com/ (visited on 23/10/2025).

[34] European Commission. *AI Pact*. Signatories of the AI Pact section. Digital Strategy, European Commission. URL: https://digital-strategy.ec.europa.eu/it/policies/ai-pact#ecl-inpage-Signatories-of-the-AI-Pact (visited on 18/11/2025).

[35] European Commission. *Building Trust in Human-Centric Artificial Intelligence*. Communication from the Commission to the European Parliament, the European Council, the Council, the European Economic and Social Committee and the Committee of the Regions COM(2019) 168 final. Communication accompanying the Ethics Guidelines for Trustworthy AI. Brussels: European Commission, Apr. 2019. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52019DC0168.

[36] European Commission. *White Paper on Artificial Intelligence: A European approach to excellence and trust*. Tech. rep. COM(2020) 65 final. Accessed: 2025-09-30. Brussels: European Commission, Feb. 2020.

[37] European Data Protection Board. *Report of the work undertaken by the ChatGPT Taskforce*. Tech. rep. European Data Protection Board, May 2024. URL: https://www.edpb.europa.eu/system/files/2024-05/edpb_20240523_report_chatgpt_taskforce_en.pdf (visited on 23/10/2025).

[38] European Parliament. *Civil Law Rules on Robotics. European Parliament resolution of 16 February 2017 with recommendations to the Commission on Civil Law Rules on Robotics*. Resolution. Procedure: 2015/2103(INL). 16th Feb. 2017. URL: https://www.europarl.europa.eu/doceo/document/TA-8-2017-0051_EN.html (visited on 27/10/2023).

[39] European Parliament. *Minutes – Results of Roll-Call Votes, 13 March 2024: Artificial Intelligence Act (A9-0188/2023) / Other Votes*. https://www.europarl.europa.eu/doceo/document/PV-9-2024-03-13-RCV_EN.html#166051#941757. Accessed: 2025-09-30. 2024.

[40] European Parliament. *Procedure file — 2021/0106(COD) Artificial Intelligence Act*. https://oeil.secure.europarl.europa.eu/oeil/en/procedure-file?reference=2021/0106(COD). Accessed: 2025-09-30. 2024.

[41] European Parliament. *Regulation on Artificial Intelligence – Legislative Train*. https://www.europarl.europa.eu/legislative-train/theme-a-europe-fit-for-the-digital-age/file-regulation-on-artificial-intelligence?sid=8801. Accessed: 2025-09-30. 2025.

[42] European Parliament. *Report on intellectual property rights for the development of artificial intelligence technologies (A9-0176/2020)*. Tech. rep. A9-0176/2020. Accessed: 2025-09-30. European Parliament, 2020.

[43] European Parliament. *Report with recommendations to the Commission on a civil liability regime for artificial intelligence (A9-0178/2020)*. Tech. rep. A9-0178/2020. Accessed: 2025-09-30. European Parliament, 2020.

[44] European Parliament. *Resolution of 20 October 2020 with recommendations to the Commission on a framework of ethical aspects of artificial intelligence, robotics and related technologies (2020/2012(INL))*. https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52020IP0275. Official Journal C 404, 6 October 2021, pp. 63–106. Accessed: 2025-09-30. 2021.

[45] Farbin Fayza et al. "Photonics for sustainable AI". In: *Communications Physics* 8.1 (Oct. 2025). ISSN: 2399-3650. DOI: 10.1038/s42005-025-02300-0. URL: http://dx.doi.org/10.1038/s42005-025-02300-0.

[46] Mark Fenwick, Wulf A. Kaal and Erik P. M. Vermeulen. "Regulation Tomorrow: What Happens When Technology Is Faster Than the Law?" In: *SSRN Electronic Journal* (2016). ISSN: 1556-5068. DOI: 10.2139/ssrn.2834531. URL: http://dx.doi.org/10.2139/ssrn.2834531.

[47] Adrian Friday et al. *The belief in Moore's Law is undermining ICT climate action*. 2024. DOI: 10.48550/ARXIV.2411.17391. URL: https://arxiv.org/abs/2411.17391.

[48] Jack Gallifant et al. "Peer review of GPT-4 technical report and systems card". In: *PLOS Digital Health* 3.1 (Jan. 2024). Ed. by Imon Banerjee, e0000417. ISSN: 2767-3170. DOI: 10.1371/journal.pdig.0000417. URL: http://dx.doi.org/10.1371/journal.pdig.0000417.

[49] Nathan Genicot. *From Blueprint to Reality: Implementing AI Regulatory Sandboxes under the AI Act*. Tech. rep. A slightly revised version of the original report published December 2024. Brussels: FARI and LSTS Research Group (VUB), 2024. URL: https://content.fari.brussels/media/f2f3eeeee3d2d52edbff1afb-genicotnimplementingairegulatorysandboxesjun25.pdf.

[50] *Global Data Broker Market: Industry Analysis and Forecast (2025-2032). Data Broker Market is expected to grow CAGR of 7.25%*. Market research report 55670. Accessed: 2025-10-19. Pune, India: Maximize Market Research, Jan. 2025. URL: https://www.maximizemarketresearch.com/market-report/global-data-broker-market/55670/.

[51] Roland Berger GmbH. *European AI Act: Opportunities and Challenges*. Insight / Publication. Roland Berger, 2025. URL: https://www.rolandberger.com/en/Insights/Publications/European-AI-Act-Opportunities-and-challenges.html (visited on 23/10/2025).

[52] Delaram Golpayegani et al. "AI Cards: Towards an Applied Framework for Machine-Readable AI and Risk Documentation Inspired by the EU AI Act". In: *Privacy Technologies and Policy*. Springer Nature Switzerland, 2024, pp. 48–72. ISBN: 9783031680243. DOI: 10.1007/978-3-031-68024-3_3. URL: http://dx.doi.org/10.1007/978-3-031-68024-3_3.

[53] Mélanie Gornet. "The AI Act: the evolution of "trustworthy AI" from policy documents to mandatory regulation". working paper or preprint. Nov. 2024. URL: https://hal.science/hal-04785519.

[54] Shweta Goyal and Sugam Gupta. "A comprehensive review of current techniques, issues, and technological advancements in sustainable E-waste management". In: *e-Prime - Advances in Electrical Engineering, Electronics and Energy* 9 (Sept. 2024), p. 100702. ISSN: 2772-6711. DOI: 10.1016/j.prime.2024.100702. URL: http://dx.doi.org/10.1016/j.prime.2024.100702.

[55] Group-IB. *The Anatomy of a Deepfake Voice Phishing Attack: How AI Voice Cloning Exploits Trust, Drains Millions, and How to Detect and Stop It*. Blog post. Aug. 2025. URL: https://www.group-ib.com/blog/voice-deepfake-scams/.

[56] Jeroen van der Heijden. "Risk Governance and Risk-Based Regulation: A Review of the International Academic Literature". In: *SSRN Electronic Journal* (2019). ISSN: 1556-5068. DOI: 10.2139/ssrn.3406998. URL: http://dx.doi.org/10.2139/ssrn.3406998.

[57] High-Level Expert Group on Artificial Intelligence. *Ethics Guidelines for Trustworthy AI*. Report. Document made public on 8 April 2019. Coordinator: Nathalie Smuha. Brussels: European Commission, Apr. 2019. URL: https://ec.europa.eu/digital-strategy/en/library/ethics-guidelines-trustworthy-ai.

[58] International Energy Agency. *Electricity 2024 – Analysis and forecast to 2026*. Revised version, January and May 2024. Licence: CC BY 4.0. Paris: International Energy Agency, 24th Jan. 2024. URL: https://iea.blob.core.windows.net/assets/18f3ed24-4b26-4c83-a3d2-8a1be51c8cc8/Electricity2024-Analysisandforecastto2026.pdf.

[59] Vijaya Krishna Kanaparthi. "Examining the Plausible Applications of Artificial Intelligence & Machine Learning in Accounts Payable Improvement". In: *FinTech* 2.3 (July 2023), pp. 1–14. DOI: None. URL: https://ideas.repec.org/a/gam/jfinte/v2y2023i3p26-474d1193090.html.

[60] Sayash Kapoor et al. *On the Societal Impact of Open Foundation Models*. 2024. arXiv: 2403.07918 [cs.CY]. URL: https://arxiv.org/abs/2403.07918.

[61] Stephanie Kirmer. *Environmental Implications of the AI Boom*. Accessed: 2025-11-09. 2nd May 2024. URL: https://medium.com/data-science/environmental-implications-of-the-ai-boom-279300a24184.

[62] Klover.ai. *Why 'Open Source' AI Isn't Always Open - and What Researchers Are Saying*. Accessed: 2025-10-16. 2024. URL: https://www.klover.ai/why-open-source-ai-isnt-always-open-and-what-researchers-are-saying/.

[63] Santeri Koivula. *AI Regulatory Sandbox Approaches: EU Member State Overview*. Accessed: 2025-11-05. May 2025. URL: https://artificialintelligenceact.eu/ai-regulatory-sandbox-approaches-eu-member-state-overview/.

[64] Martin Kretschmer et al. "The risks of risk-based AI regulation: taking liability seriously". In: *SSRN Electronic Journal* (2023). ISSN: 1556-5068. DOI: 10.2139/ssrn.4622405. URL: http://dx.doi.org/10.2139/ssrn.4622405.

[65] Anuj Kumar, Anirudh Thorbole and Ram K. Gupta. "Sustaining the future: Semiconductor materials and their recovery". In: *Materials Science in Semiconductor Processing* 185 (Jan. 2025), p. 108943. ISSN: 1369-8001. DOI: 10.1016/j.mssp.2024.108943. URL: http://dx.doi.org/10.1016/j.mssp.2024.108943.

[66] Andreas Liesenfeld and Mark Dingemanse. "Rethinking open source generative AI: open washing and the EU AI Act". In: *The 2024 ACM Conference on Fairness, Accountability, and Transparency*. FAccT '24. ACM, June 2024, pp. 1774–1787. DOI: 10.1145/3630106.3659005. URL: http://dx.doi.org/10.1145/3630106.3659005.

[67] *Ligue des droits humains ASBL v Conseil des ministres*. Judgment of the Grand Chamber (Request for a preliminary ruling). 21st June 2022. URL: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:62019CJ0817.

[68] Amy Luers et al. "Will AI accelerate or delay the race to net-zero emissions?" In: *Nature* 628 (22nd Apr. 2024), pp. 718–720. DOI: 10.1038/d41586-024-01137-x. URL: https://www.nature.com/articles/d41586-024-01137-x.

[69] Carl Macrae and Christopher K. Ansell. "Generative Spaces: Collaboration, Learning and Innovation in a Regulatory Sandbox". In: *SSRN Electronic Journal* (2024). ISSN: 1556-5068. DOI: 10.2139/ssrn.4825907. URL: http://dx.doi.org/10.2139/ssrn.4825907.

[70] Christina Majaski. *Understanding Rent Seeking: Economics Definition & Examples*. Investopedia. 4th Oct. 2025. URL: https://www.investopedia.com/terms/r/rentseeking.asp (visited on 07/11/2025).

[71] Avinash Malladhi. "From Manual to Automated: The Transformation of Accounting Information Systems Through AI-OCR Technology". In: *TIJER - International Research Journal* 10.7 (2023). TIJER2307017, July 2023, pp. 130–137. ISSN: 2349-9249. URL: https://www.tijer.org.

[72] Bertin Martens. *Catch-up with the US or prosper below the tech frontier? An EU artificial intelligence strategy*. Bruegel Policy Brief No. 25/2024. Bruegel, 2024. URL: https://www.bruegel.org/policy-brief/catch-us-or-prosper-below-tech-frontier-eu-artificial-intelligence-strategy.

[73] Tshilidzi Marwala. *Rethinking Tech and Why GPUs Are Not the Future of AI Training*. Accessed: 2025-11-09. 2025. URL: https://unu.edu/article/rethinking-tech-and-why-gpus-are-not-future-ai-training.

[74] Joanna Mazur and Marcin Serafin. "Stalling the State: How Digital Platforms Contribute to and Profit From Delays in the Enforcement and Adoption of Regulations". In: *Comparative Political Studies* 56.1 (June 2022), pp. 101–130. ISSN: 1552-3829. DOI: 10.1177/00104140221089651. URL: http://dx.doi.org/10.1177/00104140221089651.

[75] Meta. *Llama FAQs — Restriction on Llama Multimodal Models in the EU*. Accessed: 20 Nov 2025. 2025. URL: https://www.llama.com/faq/%5C#Restriction%5C%20on%5C%20Llama%5C%20Multimodal%5C%20Models%5C%20in%5C%20the%5C%20EU.

[76] L. Mozgunova. "A Critical Overview of the Fundamental Aspects of the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence". en. In: *SSRN* (2024). URL: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4724557.

[77] Fabian Muhly, Emanuele Chizzonic and Philipp Leo. "AI-deepfake scams and the importance of a holistic communication security strategy". In: *International Cybersecurity Law Review* 6.1 (Feb. 2025), pp. 53–61. ISSN: 2662-9739. DOI: 10.1365/s43439-025-00143-7. URL: http://dx.doi.org/10.1365/s43439-025-00143-7.

[78] National Counterterrorism Innovation, Technology, and Education Center (NCITE). *Deepfakes and Fraud: Real-World Examples of AI Misuse*. Report 136. University of Nebraska at Omaha, Dr. C.C. and Mabel L. Criss Library, June 2025. URL: https://digitalcommons.unomaha.edu/nciterepportsresearch/136.

[79] Claudio Novelli et al. "AI Risk Assessment: A Scenario-Based, Proportional Methodology for the AI Act". In: *Digital Society* 3.1 (Mar. 2024). ISSN: 2731-4669. DOI: 10.1007/s44206-024-00095-1. URL: http://dx.doi.org/10.1007/s44206-024-00095-1.

[80] Cathy O'Neil. *Weapons of math destruction*. Crown Publishing Group, Sept. 2016.

[81] U. Okoro. "Artificial Intelligence Governance and Regulation; The impact of the EU AI Act, 2024 on Innovation, Accountability, and Global Compliance in a Digital Age". en. In: *SSRN* (2025). DOI: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5138252.

[82] Open Source Initiative. *Final Board Report: The Open Source AI Definition*. https://opensource.org/ai/final-board-report. Accessed: 2025-10-04. 2025.

[83] Open Source Initiative. *The Open Source AI Definition – 1.0*. https://opensource.org/ai/open-source-ai-definition. Accessed: 2025-10-04. 2025.

[84] OpenAI. *Introducing GPT-OSS*. Accessed: 2025-10-16. URL: https://openai.com/index/introducing-gpt-oss/.

[85] OpenAI. *Models Documentation*. Accessed: 2025-10-16. URL: https://platform.openai.com/docs/models.

[86] OpenAI. *Open Models*. Accessed: 2025-10-16. URL: https://openai.com/open-models/.

[87] OpenAI. *OpenAI Open-Weight Models (GPT-OSS)*. Accessed: 2025-10-16. URL: https://help.openai.com/en/articles/11870455-openai-open-weight-models-gpt-oss.

[88] OpenAI. *OpenAI Open-Weight Models (gpt-oss)*. Accesso: 16 ottobre 2025. 2025. URL: https://help.openai.com/en/articles/11870455-openai-open-weight-models-gpt-oss.

[89] *OpenAI GPT-4*. Company Report Card. Stanford University Center for Research on Foundation Models (CRFM), Foundation Model Transparency Index (FMTI). May 2024. URL: https://crfm.stanford.edu/fmti/May-2024/company-reports/OpenAI_GPT-4.html (visited on 18/11/2025).

[90] *OpenAI, Microsoft, Amazon Among First AI Pact Signatories*. Euronews Next. 25th Sept. 2024. URL: https://www.euronews.com/next/2024/09/25/openai-microsoft-amazon-among-first-ai-pact-signatories (visited on 18/11/2025).

[91] *OpenAI: New noyb GDPR Complaint in EU*. Digital Market Reports. 1st Dec. 2023. URL: https://digitalmarketreports.com/news/15794/openai-new-noyb-gdpr-complaint-in-eu/ (visited on 18/11/2025).

[92] Carsten Orwat et al. "Normative Challenges of Risk Regulation of Artificial Intelligence". In: *NanoEthics* 18.2 (Aug. 2024). ISSN: 1871-4765. DOI: 10.1007/s11569-024-00454-9. URL: http://dx.doi.org/10.1007/s11569-024-00454-9.

[93] James Ostrowski. "Regulating Machine Learning Open-Source Software". In: *Abundance Institute* (May 2024). Accesso: 16 ottobre 2025. URL: https://abundance.institute/articles/regulating-machine-learning-open-source-software.

[94] Michael Park, Shuping Wu and Russell J. Funk. "Regulation and Innovation Revisited: How Restrictive Environments Can Promote Destabilizing New Technologies". In: *Organization Science* 36.1 (Jan. 2025), pp. 240–260. ISSN: 1526-5455. DOI: 10.1287/orsc.2022.16770. URL: http://dx.doi.org/10.1287/orsc.2022.16770.

[95] José Renato Laranjeira de Pereira. *The EU AI Act and environmental protection: the case for a missed opportunity*. Heinrich-Böll-Stiftung. 8th Apr. 2024. URL: https://us.boell.org/en/2024/04/08/eu-ai-act-and-environmental-protection-case-missed-opportunity.

[96] Garante per la protezione dei dati personali. *Intelligenza artificiale: il Garante blocca ChatGPT. Raccolta illecita di dati personali. Assenza di sistemi per la verifica dell'età dei minori*. Comunicato stampa Doc-Web 9870847. Roma: Garante per la protezione dei dati personali, Mar. 2023. URL: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/9870847 (visited on 23/10/2025).

[97] Garante per la protezione dei dati personali. *Provvedimento del 30 gennaio 2025: Limitazione definitiva del trattamento dei dati personali da parte delle società Hangzhou DeepSeek Artificial Intelligence Co., Ltd. e Beijing DeepSeek Artificial Intelligence Co., Ltd.* Provvedimento Doc-Web 10098477. Registro dei provvedimenti n. 33 del 30 gennaio 2025. Roma: Garante per la protezione dei dati personali, Jan. 2025. URL: https://www.garanteprivacy.it/web/guest/home/docweb/-/docweb-display/docweb/10098477 (visited on 23/10/2025).

[98] Politico Staff. *Europe's privacy patrol is out for vengeance to block AI*. URL: https://www.politico.eu/article/europe-privacy-patrol-vengeance-block-ai-artificial-intelligence/ (visited on 23/10/2025).

[99] Chris Pooley and Viviana Gomes. *How FibroGen achieved 40x ROI by automating invoice processing*. Accessed: 2025-10-27. Google Cloud. 13th Mar. 2024. URL: https://cloud.google.com/blog/products/ai-machine-learning/reducing-invoice-processing-with-document-ai/.

[100] Richard Pope. *Platformland*. en. London, England: London Publishing Partnership, Sept. 2024.

[101] Michael E. Porter and Claas van der Linde. "Toward a New Conception of the Environment-Competitiveness Relationship". In: *Journal of Economic Perspectives* 9.4 (Dec. 1995), pp. 97–118. DOI: 10.1257/jep.9.4.97. URL: https://www.aeaweb.org/articles?id=10.1257/jep.9.4.97.

[102] *Preliminary noyb Win: Meta Stops AI Plans in EU*. noyb European Center for Digital Rights. 17th June 2024. URL: https://noyb.eu/en/preliminary-noyb-win-meta-stops-ai-plans-eu (visited on 18/11/2025).

[103] *Privacy Technologies and Policy: 12th Annual Privacy Forum, APF 2024, Karlstad, Sweden, September 4–5, 2024, Proceedings*. Springer Nature Switzerland, 2024. ISBN: 9783031680243. DOI: 10.1007/978-3-031-68024-3. URL: http://dx.doi.org/10.1007/978-3-031-68024-3.

[104] Sofia Ranchordas and Valeria Vinci. "Regulatory Sandboxes and Innovation-friendly Regulation: Between Collaboration and Capture". In: *SSRN Electronic Journal* (2024). ISSN: 1556-5068. DOI: 10.2139/ssrn.4696442. URL: http://dx.doi.org/10.2139/ssrn.4696442.

[105] Sebastian Raschka. *Build a large language model (from scratch)*. en. New York, NY: Manning Publications, Nov. 2024.

[106] Tytti Rintamäki et al. "Impact Assessment Requirements in the GDPR vs the AI Act: Overlaps, Divergence, and Implications". In: (May 2025). DOI: 10.31219/osf.io/6qhzj_v2. URL: http://dx.doi.org/10.31219/osf.io/6qhzj_v2.

[107] *Risk and Regulatory Policy: Improving the Governance of Risk*. OECD, Apr. 2010. ISBN: 9789264082939. DOI: 10.1787/9789264082939-en. URL: http://dx.doi.org/10.1787/9789264082939-en.

[108] Adriana Robertson. "Timing the Regulatory Tightrope". In: *SSRN Electronic Journal* (2023). ISSN: 1556-5068. DOI: 10.2139/ssrn.4593670. URL: http://dx.doi.org/10.2139/ssrn.4593670.

[109] Stuart Russell and Peter Norvig. *Artificial intelligence*. en. 4th ed. Upper Saddle River, NJ: Pearson, June 2021.

[110] Stuart Russell, Karine Perset and Marko Grobelnik. *Updates to the OECD's definition of an AI system explained*. https://oecd.ai/en/wonk/ai-system-definition-update. Accessed: 2025-09-30. OECD AI, Nov. 2023.

[111] Ranjan Sapkota, Shaina Raza and Manoj Karkee. *Comprehensive Analysis of Transparency and Accessibility of ChatGPT, DeepSeek, And other SoTA Large Language Models*. 2025. arXiv: 2502.18505 [cs.SE]. URL: https://arxiv.org/abs/2502.18505.

[112] Nuno Sousa e Silva. *The Artificial Intelligence Act: critical overview*. 2024. DOI: 10.48550/ARXIV.2409.00264. URL: https://arxiv.org/abs/2409.00264.

[113] Josh Sisco. *The New York Times Sues OpenAI, Microsoft Over Use of Its Stories Used to Train Chatbots*. Politico. 27th Dec. 2023. URL: https://www.politico.com/news/2023/12/27/the-new-york-times-sues-openai-microsoft-over-use-of-its-stories-used-to-train-chatbots-00133234 (visited on 18/11/2025).

[114] Eleni Skanavi. *Max Schrems Takes on ChatGPT: Can AI Be Made GDPR Compliant?* Maastricht Student Law Review. 20th Nov. 2023. URL: https://www.maastrichtstudentlawreview.com/post/max-schrems-takes-on-chatgpt-can-ai-be-made-gdpr-compliant (visited on 18/11/2025).

[115] Kasia Söderlund and Stefan Larsson. "Enforcement Design Patterns in EU Law: An Analysis of the AI Act". en. In: *Digital Society* 3.2 (2024). DOI: 10.1007/s44206-024-00129-8. URL: http://dx.doi.org/10.1007/s44206-024-00129-8.

[116] Tarun Tater et al. "AI Driven Accounts Payable Transformation". In: *Proceedings of the Thirty-Sixth AAAI Conference on Artificial Intelligence (AAAI-22)*. Accessed: 2025-10-27. Palo Alto, CA: Association for the Advancement of Artificial Intelligence, 2022. URL: https://www.academia.edu/91435753/AI_Driven_Accounts_Payable_Transformation.

[117] Max von Thun. *To Innovate or to Regulate? The False Dichotomy at the Heart of Europe's Industrial Approach*. Research Publication. Accessed: 23 October 2025. AI Now Institute, Mar. 2024. URL: https://ainowinstitute.org/publications/to-innovate-or-to-regulate-the-false-dichotomy (visited on 23/10/2025).

[118] Super User. *Comparative Analysis of Mistral 7B and OLMo2 7B on the Ollama Platform*. Accessed: 2025-10-16. Mar. 2025. URL: https://matasoft.hr/qtrendcontrol/index.php/un-perplexed-spready/un-perplexed-spready-various-articles/146-comparative-analysis-of-mistral-7b-and-olmo2-7b-on-the-ollama-platform.

[119] Roberto Verdecchia, June Sallou and Luís Cruz. *A Systematic Review of Green AI*. 2023. arXiv: 2301.11047 [cs.AI]. URL: https://arxiv.org/abs/2301.11047.

[120] Nicholas Vinocur. "One country blocks the world on data privacy". In: *Politico Europe* (Apr. 2019). URL: https://www.politico.eu/interactive/ireland-blocks-the-world-on-data-privacy/ (visited on 23/10/2025).

[121] Andy Z. Wang. "Network Harms". In: *University of Chicago Law Review* 91.7 (2024). Accessed: 2025-10-16.

[122] Wayne Wei Wang et al. "Artificial Intelligence "Law(s)" in China: Retrospect and Prospect". In: (2024). DOI: 10.2139/ssrn.5039316. URL: http://dx.doi.org/10.2139/ssrn.5039316.

[123] Matthew T. Wansley. "Regulation of Emerging Risks". In: *Vanderbilt Law Review* 69.2 (2016). Accessed: 2025-11-10, pp. 401–. URL: https://scholarship.law.vanderbilt.edu/vlr/vol69/iss2/3/.

[124] Matt White et al. *The Model Openness Framework: Promoting Completeness and Openness for Reproducibility, Transparency, and Usability in Artificial Intelligence*. 2024. eprint: arXiv:2403.13784.

[125] *Why OpenAI's Voice Mode, Meta's Llama and Apple's AI Won't be Coming to Europe Yet*. Euronews Next. 8th Oct. 2024. URL: https://www.euronews.com/next/2024/10/08/why-openais-voice-mode-metas-llama-and-apples-ai-wont-be-coming-to-europe-yet (visited on 18/11/2025).

[126] ZDNet. "The best open source AI models: all your free-to-use options explained". In: (2025). Accessed: 2025-10-04.

[127] Angela Huyue Zhang. "The Promise and Perils of China's Regulation of Artificial Intelligence". In: *SSRN Electronic Journal* (2024). ISSN: 1556-5068. DOI: 10.2139/ssrn.4708676. URL: http://dx.doi.org/10.2139/ssrn.4708676.

[128] Xin Zhang et al. "How to monetize data: An economic analysis of data monetization strategies under competition". In: *Decision Support Systems* 173 (Oct. 2023). ISSN: 0167-9236. DOI: 10.1016/j.dss.2023.114012. URL: http://dx.doi.org/10.1016/j.dss.2023.114012.

[129] 清大智慧法治研究院. *FTC如何管据人：以X-Mode案例*. https://www.secrss.com/articles/65200. Accessed: 2025-10-17. Apr. 2024.

Part III

Remaining Legal and Practical Challenges

Chapter 12

The AI Act - GDPR interplay

Compliance with the Regulation does not automatically mean that a model's use is *lawful* under other Union or national laws. The use of any AI system must comply with *all* relevant legal requirements, including those preceding the AI Act — especially the GDPR with which it is **complementary**: the GDPR is a principles-based, technology-neutral law aimed at safeguarding personal data, while the AI Act introduces a risk-based approach to regulating AI applications that use said data.

There is even significant *overlap* between the two laws, as one is considered a *lex specialis* for the other. For example, the AI Act's rules on data quality and bias closely align with the GDPR's principles of *fairness* and *data accuracy*. The AI Act's requirements for risk management systems, human oversight, and transparency reinforce GDPR obligations, particularly for issues such as data subject rights (access, rectification, erasure). Similarly, the need for a Data Protection Impact Assessment under the GDPR is reflected by the AI Act's requirement for a fundamental rights impact assessment for high-risk systems.

Despite the AI Act's comprehensiveness, some issues still fall through the cracks between laws, and some challenges remain unresolved.

12.1 Is the Model Itself Personal Data? [19, 6, 9]

Whether a Large Language Model itself qualifies as «*personal data*» remains a highly debated legal question, with no clear consensus and compelling arguments on both sides.

The Argument for LLMs as Personal Data Legal scholars and researchers have argued that once an LLM memorises and is able to reproduce personal information from its training data, **the model itself becomes personal data**. This view aligns with the GDPR's intentionally broad definition of personal data, which covers any information relating to an identified or identifiable individual. Its technologically neutral approach is particularly fitting for LLMs, as recent research shows that, with carefully crafted prompts, it is possible to extract personal information from these models, reinforcing the idea that an LLM can be a repository of personal data.There is broad agreement that the format in which information is stored does not change this reality: even if LLMs encode data as abstract parameters or weights that are not immediately interpretable, it is still information relating to an identifiable person.

The implications of this view are far-reaching. If an LLM is recognised as personal data, anyone who trains, shares, or deploys the model in Europe must comply with significant legal obligations. This includes securing a lawful basis for processing and protecting data subject rights, such as the right to access, rectify, or erase personal data directly from the model.

The Alternative View: Focus on Outputs, Not the Model Some authorities, such as Hamburg's DPA, take a different view. They argue that the parameters and weights of an LLM should not be classified as personal data because they function as *probabilistic systems* that express statistical patterns from large datasets and produce impersonal «summaries», represented mathematically as probabilities.

This approach prefers to shift the regulatory focus to the inputs and outputs of the AI system. While the internal workings of the model may remain opaque, the GDPR's protections would only apply to personal data that appears in system outputs. This is relatively easy to manage, as developers can filter outputs after the fact rather than undertaking the unrealistic and resource-intensive task of retraining the entire model when a breach occurs.

12.2 Mapping GDPR Principles against the reality of an AI model

The GDPR was created at a time when personal data was typically stored in traditional databases, organised neatly in tables and forms. Large Language Models, however, operate in a fundamentally different way. During training, LLMs absorb vast amounts of data and encode it as complex patterns distributed across billions of parameters. As a result, it becomes **nearly impossible to alter or erase an individual data point after training** without damaging the rest. This fundamental difference creates friction between the GDPR and LLMs, especially when it comes to transparency and accountability, and the rights to be forgotten and to correct inaccurate personal information, which risk remaining only on paper.

Addressing these challenges will require **innovation in technology, governance, and regulatory interpretation**. Researchers are developing *privacy-preserving machine learning* techniques, using tools such as **differential privacy, federated learning, and data anonymisation**, **to minimise the trade-off between model utility and privacy**. *Explainable AI* (*XAI*) is also being explored to improve transparency and auditability. From a governance perspective, implementing *Privacy by Design*, as required by *Article 25* of the GDPR, is essential. This means **embedding data protection considerations throughout the LLM development lifecycle**, rather than treating them as an afterthought. Comprehensive governance frameworks have been proposed to address this task through data governance, consent management, and ongoing monitoring.

As of today, AI,and LLMs in particular,continues to test the boundaries of data protection law. The following section reviews new safeguards and strategies intended to address these ongoing challenges.

12.2.1 The Right to Erasure and the role of Machine Unlearning [34, 14, 10, 15, 17, 19, 24, 4]

The *Right to Erasure*, also known as the *Right to be Forgotten*, provides individuals with a way to challenge the widespread belief that whatever is put on the Internet stays there forever. *Article 17* of the GDPR allows people to request the **deletion** of their personal data if keeping it would harm their rights and is not justified by reasons such as public interest. However, Large Language Models have made enforcing this right extremely difficult. Unlike deleting a record from a conventional database, **making a model forget a single data point is extremely challenging from both a technical and practical perspective**, as information is deeply intertwined and represented as abstract parameters, so attempting to remove one piece of information can disrupt much more than just a single data point. Moreover, the training process itself is complex, automated, dynamic, and unpredictable, so it is nearly impossible to know in advance how any single data point will affect the final model. Models may retain subtle traces of that information in ways that are difficult to detect, and the same applies to any derivative models. Retraining a model from scratch is not a viable solution either, as it is extremely expensive and time-consuming. For companies that have invested months and millions of dollars in training, retraining because of a single user's request is simply not feasible.

Additionally, simply deleting a record from the training set is not sufficient if models trained on earlier versions still exist. Additionally, Because of these factors, the right to erasure under the GDPR is often impossible to achieve in practice when it comes to Large Language Models, for both technical and economic reasons.

12.2.1.1 The role of Machine Unlearning

Machine Unlearning has emerged as a critical field of research on how to «scrub away» specific data points from a trained model, **yielding a result indistinguishable from a model that was never trained on that data to begin with**.

Machine Unlearning can be divided into two broad categories, with the first including the so-called *exact unlearning*, which is, as of today, the most straightforward, but computationally burdensome, solution: in this approach, the entire model is **retrained from scratch** on a dataset that omits the data subject to erasure, thereby offering a theoretically proven guarantee that all influence of the targeted data will be eliminated. Such a guarantee is obviously the preferred one from the perspective of GDPR compliance [14], but it obviously requires enormous computational resources and time, so it is not viable for real-world, large-scale scenarios.

In contrast, the second category, also known as *approximate unlearning*, is much more efficient and functions at a fraction of the cost, which makes it the preferred way despite its lack of the same theoretical certainty (the original influence of the target data might still linger slightly) [25]. It *approximates* the state it would have been in had it not been trained on the target data. Techniques within this category include, for example: **gradient-based updates**, in which a gradient ascent is used to «reverse» the learning just as a gradient descent was used to learn originally, «*intentionally maximis[ing] the loss for those specific samples*» [16]; **data partitioning** such as the *SISA* model[1] [14],

[1] Sharded, Isolated, Sliced, and Aggregated.

in which the training dataset is distributed into isolated pieces, each used to train a part of the model and thus, when a data point needs to be forgotten, only the relevant submodel needs to be retrained [4]; and **influence functions**, used to estimate the effect of a data point on the model's parameters and predictions, to counteract it with targeted fine-tuning [9].

12.2.1.2 Adequateness of the approach [14]

A key question is whether these unlearning methods are **adequate to meet the legal requirements** of *Article 17*. The GDPR does allow for some flexibility, requiring only *reasonable* steps that take into account current technology and the cost of implementation. The interpretation of this «*reasonableness*» requirement is central to the legal debate and depends on the specific risks in each situation; if a model is shown to leak identifiable personal data, then only the most robust and comprehensive method, namely, *exact unlearning through full retraining*, is considered to fully satisfy legal obligations. On the other hand, if there is no clear evidence that a model's outputs can directly identify individuals, approximate unlearning methods, especially when combined with *output filtering* and preventative sanitisation of the training data, may be considered a reasonable compromise. The field is still developing quickly, and it remains to be seen how effective or costly machine unlearning techniques will become.

12.2.2 The Right of Access in an Opaque System [9, 34, 19, 6]

The *Right of Access*, set out in Article 15 of the GDPR, allows individuals to find out what personal data is held about them, how it is processed, and what types of data are involved. This right is a cornerstone of the GDPR, and it is essential for individuals to exercise also other rights, such as requesting the correction or deletion of their information.

It is not surprising that enforcing this right presents significant technical challenges for Large Language Models. The main difficulties arise from the way data is deeply intertwined within the model, as well as the so-called black-box nature of these systems. As a result, most models do not provide a straightforward way for users to access or retrieve their personal information. Depending on the case, this may be due to technical limitations, but also to lack of initiative from developers, or simple oversight. Attempts to query the model directly to determine its contents have proven unreliable, as the model may hallucinate responses, fail to recall actual memorised data, or be unable to reconstruct the reasoning behind its answers.

To address these challenges, researchers and organisations are developing a range of solutions. One approach is to create **specialised model auditing tools** that can detect whether personal data is present within a model's parameters. Another strategy is to embed **robust data provenance tracking into the development pipeline**, so that every piece of training data can be traced from its origin to its use in the model. Other measures include allowing users to delete conversation logs from chat-based services or to use private sessions, where shared information is not used for model training and only temporary logs are kept for security purposes.

At the governance level, some companies have begun publishing **transparency reports** about their data sources. These reports can give users a general sense of whether their information may have been included in training, even if specifics are not provided.

12.2.3 AI Hallucinations and the GDPR's Accuracy Principle[6, 13, 8, 19, 9, 34]

AI hallucinations **Hallucinations** occur when AI models generate responses that sound convincing but are in fact false or nonsensical. This happens because Large Language Models are **designed to predict the next most likely word based on patterns in their training data, rather than to check facts, and** were not intended to link their answers to external, verifiable sources (although recent developments have made some progress in this area). Their fluency comes from understanding language structure, rather than truly grasping the content.

The inability to distinguish between statistical associations and actual facts can cause significant problems in sensitive fields such as finance and law. Hallucinations may even contradict the original meaning of the source material, often due to issues with the model's attention mechanism, which is essential for tracking relationships in data over long passages. At times, these errors introduce information that cannot be verified or was never present in the training input.

AI hallucinations have various **causes**, which may arise at several stages of a model's lifecycle, from data collection through to training and inference. One major factor is **flawed training data**: if the datasets contain errors, biases, or outdated information, the model learns patterns that are not accurate. **Architectural and training choices** also play a significant role. For instance, **exposure bias** can develop when the model's training data does not match what it encounters during real-world use. Techniques such as *Reinforcement Learning from Human Feedback* can sometimes encourage the model to produce answers that will please human evaluators, even if those answers are less accurate. Lastly, hallucinations may also occur at the **inference stage**: when randomness is introduced to create more varied responses, it increases the chance of errors, and when the model is asked for very lengthy outputs, its attention mechanism may begin to lose track of the original context, a problem known as **context hijacking**.

Figure 12.1: By Randall Munroe. Creative Commons Attribution-NonCommercial 2.5 License.

The role of European Law The AI Act addresses hallucinations by requiring **high standards for accuracy and robustness**. These requirements are meant to reduce the risks and harms caused by AI-generated errors.

Despite the efforts, hallucinations remain a major challenge, especially in light of the GDPR's principle of *accuracy*, which is violated every time an AI system generates

incorrect personal information.

So far, the main solution has been to **encourage users to think critically about what AI models produce**. Many systems now include warnings that the outputs may be inaccurate, and they ask users to provide feedback when errors occur. Another approach is to have the model perform **web searches** to check its answers. While this can help, it also introduces new risks, since the Internet itself is a frequent source of misinformation.

The current situation Several high-profile complaints regarding hallucinations have brought this issue to public attention. For example, in April 2024, the digital rights group Noyb filed a complaint with the Austrian DPA, claiming that ChatGPT violated the accuracy principle by giving a public figure's incorrect date of birth. In another case from March 2023, ChatGPT mistakenly accused a user of being a child murderer [20]. Since European data protection authorities are still reviewing these matters and have not issued a final decision, it appears that Big Tech companies currently have significant responsibility for handling these problems in good faith.

12.2.4 Automated Judgements and Inherited Bias: Upholding Human Dignity under Article 22[9, 35, 28, 13, 8]

Automated decisions made by complex algorithms, with no possibility of appeal or redress and often trained on potentially biased data, pose a genuine **risk to human dignity**. When individuals are treated as mere data points and are powerless in the face of decisions that affect them, it undermines their freedom and the opportunity for fair treatment in society.

Article 22 of the GDPR aims to protect individuals from decisions made solely by automated means, especially when these decisions significantly affect their lives. The provision grants people the right to request a human review, present their perspective, and have decisions reconsidered. However, modern AI complicates this process, as seeking human intervention may be seen as inefficient and cumbersome, particularly when these tools are used with large populations. In contrast, traditional decision-making methods continue for smaller, typically more privileged groups. This problem impacts not only legal rights but also the individual's sense of control, which is vital to human dignity.

Rather than evaluating unique circumstances, automated decision tools in contexts like credit or recruitment neglect individuality and reject the idea that every person has intrinsic worth and must not be treated as a collection of data points.

To uphold human dignity, regulatory frameworks and ethical guidelines emphasise the need for robust and meaningful human oversight. Investing in **XAI** techniques remains one of the most promising ways to achieve this at present.

12.2.4.1 Explainable AI

Explainable AI (*XAI*) refers to a set of methods and techniques intended to make the decision-making processes of Artificial Intelligence systems **understandable** to humans, by providing **clear explanations** for their outputs and behaviours.

Several techniques are being developed to make this a concrete possibility. Some examples are:

- *Model-agnostic techniques* such as *LIME* (*Local Interpretable Model-agnostic Explanations*) and *SHAP* (*Shapley Additive exPlanations*) [9]. Unfortunately, it has been shown that they can struggle with **non-linear and distributed knowledge**, and may not capture the full **context** contributing to a particular response.

- *Transformer-Specific Methods*, tailored for transformer architectures, include, among others, *attention weight visualisation*, *attention rollout*, and *gradient-based saliency mapping* [9]. These methods attempt to **trace the model's decision-making pathways**, although studies suggest that the workings of attention mechanisms (which aim to identify and retain only the most important pieces of information to define the context) do not always correlate reliably with the model's overall reasoning process, as they are just one part of a much more complex computational system involving other non-linear components such as *Feed-Forward Networks* [17] (where a significant portion of the model's factual and associative knowledge is believed to be stored and processed) [4].

- *Mechanistic Interpretability* [13]: This emerging field promises to find a way to **reverse-engineer** the step-by-step decision algorithms learnt by complex neural networks. It showed early promise when applied to simpler models, but achieving advanced interpretability for more complex systems remains incredibly difficult and labour-intensive. Nevertheless, it is considered a valuable direction as of today.

Despite the development of various XAI techniques, **a significant explainability gap remains**, particularly with large black-box LLMs. There are currently no [13] adequate or standardised tools to rigorously measure the extent of their explainability or to compare interpretability beyond indirect metrics that do not address the broader requirement of explainability (e.g. calibration, which assesses how well a model's predicted confidence in its output matches its actual accuracy [13], providing at least an indication of whether the model found sufficient supporting data in its knowledge base to justify a confident answer). This remains a pressing issue, as it is critical for ensuring regulatory compliance and building user trust. The impetus from the EU AI Act is expected to stimulate further research and development in this area, hopefully leading to more robust and legally viable interpretability solutions.

12.3 The scope of GDPR in the Age of Cross-Border AI

A distinguishing feature of the GDPR is its *extraterritorial scope*: its provisions can extend to actors outside the EU and apply globally. This far-reaching effect is particularly noteworthy in a world where the Internet has removed physical barriers, but has taken on new significance in the context of Artificial Intelligence, as AI systems and LLMs are built from data collected indiscriminately by actors distributed worldwide—or, even more problematically, by automated tools with no clear legal residence at all.

The GDPR's territorial scope, as set out in *Article 3*, is defined by both the *establishment* criterion and the *targeting* criterion. This extends the GDPR's reach to organisations without a physical presence in the EU if they offer goods or services to individuals in the EU, or monitor their behaviour within the Union. The latter provision is especially relevant for AI providers whose products and services are accessible to European users, even if the providers are based outside the EU, which is almost always the case. These two essential elements have in the past already legitimised the intervention of European data protection authorities over social media platforms that have a significant, albeit *virtual*, presence in Europe.

Still, **enforcing** the GDPR's extraterritorial provisions presents significant challenges. This can result in incomplete solutions, such as geo-blocking, which technologically adept users can easily circumvent. Moreover, as it happened in the landmark Google v CNIL case [5] concerning the right to be forgotten, the court itself can decline to enforce an order to de-list information from all global domains outside the European territory if a satisfactory «patch» has been put on European domains for the same website.

The most effective way to address this extraterritorial issue remains for the GDPR to continue establishing itself as a global benchmark for data protection legislation, encouraging distant jurisdictions to adopt similar rules based on the same fundamental principles and, by default, to recognise one another's decisions.

12.4 Technical Challenges

Large Language Models pose **a wide range of sophisticated threats to personally identifiable information** (PII). One of the most recognised privacy risks is **the leakage of sensitive information from training data**, but, recent research indicates that the risks extend beyond just memorisation and subsequent output, as sensitive information shared by users in prompts is exposed to the service's providers. This is particularly concerning, as it is possible to **infer sensitive user attributes, such as location or health status, from seemingly innocuous user queries**.

Threats can emerge at every stage of a model's lifecycle, from pre-training and fine-tuning to deployment and there is growing awareness of more subtle risks, such as **inference attacks** and **agent-based leakage**. These threats can generate misleading model outputs, or aggregate scattered, publicly available information about an individual to build a detailed profile, which can then be used to facilitate other types of attacks, such as password retrieval or spear phishing.

12.4.1 Beyond the Black Box: The Limits of Explainability in Transformer Architectures

The issue of transparency and explainability has already been covered at 9.4, as it is a problem that encompasses both Regulations.

12.4.2 Training Data Leakages

LLM training sets contain personal, copyrighted, or even confidential information that has been collected from across the *infosphere*, aided by the phenomenon of *data brokerage*

discussed above. The likelihood that private, sensitive information will be included in output sequences increases as models grow in size and complexity, since the number of parameters—now in the billions—provides considerable capacity to store ever more detail.

The memorisation process is influenced by several factors, and **the likelihood of a piece of information being retained depends**, for example, on how often a particular pattern appears in the training data. Studies have shown that a sequence duplicated ten times may be reproduced by a model approximately 1000 times more frequently than unique sequences. The number of parameters also directly correlates with the model's ability to store and recall specific examples from its training corpus, as does the point during the training process at which a particular piece of information is encountered: models tend to «remember» better what is seen towards the end of training [18].

Below are two examples of possible attacks that leverage the model's memorisation and repetition behaviour.

Privacy Attacks Exploiting Data Leakage A determined adversary has been shown to be able to exploit *memorisation* to extract sensitive information from LLMs. Most disturbingly, this kind of attacks do not need more than the standard *black-box* access to the model, meaning it is sufficient to access a public API without needing to see any internal architecture or weights, as it is the case with LLMs today [25]. The first step is to perform a *Training Data Extraction* [16], in which the attacker designs specific prompts to induce the model to reveal the exact content in its training data. The process generally involves generating numerous text *prefixes*, and then asking the model to produce possible continuations; having gathered multiple candidates, a filtering technique is then used to identify those most likely to be the actual memorised training data. Researchers [18, 16]have demonstrated the effectiveness of these attacks on production-level models such as ChatGPT and open-source LLMs, successfully extracting personal data, copyrighted content, and other sensitive information.

Figure 12.2: By Randall Munroe. Creative Commons Attribution-NonCommercial 2.5 License.

Membership Inference Attacks (MIAs) MIAs aim to determine whether a specific piece of data was included in a model's training set [15]. These attacks operate by **observing differences in the model's behaviour**, such as higher confidence scores, when it encounters data it was trained on compared to new data. While this does not equate to a direct leakage of content, an MIA can be the crucial part of more complex data extraction attacks [17] and can still expose sensitive information; for example, simply confirming that a person's medical record is *present* in the data used for a healthcare model specialised in a certain disease is a breach of privacy in its own right.

12.4.2.1 Mitigation strategies [16, 3, 14, 19, 30, 28, 33, 34, 22, 4, 17, 12, 23, 11, 24, 1]

A multi-layered approach is required to mitigate the risks of data leakage, to address the issue at all stages of the LLM lifecycle.

Pre-Training: Data-Level Solutions The following are two examples of methods studied by academia for their potential to clean data prior to its use in training:

- **Data anonymisation and sanitisation**: the identification and removal of PII from datasets using techniques that range from simple regular expressions to more advanced, bespoke tools. Unfortunately, these methods often lack formal privacy guarantees [11].
- **Data De-duplication** [18]: Removing duplicate sequences from training data is an effective way to reduce the risk of memorisation for some disproportionately represented information. Although it is neither a comprehensive nor a guaranteed solution, it can help mitigate the risk of leaking some highly repeated information [25].

During Training: Model-Level Solutions These techniques integrate privacy protections into the training process itself.

- **Differential Privacy (DP)**: DP adds mathematically calibrated noise during training to ensure that no single data point can be singled out as belonging to the set. DP is a tried and tested method already widely applied in GDPR compliance, as it offers **quantitative, mathematically proven guarantees of privacy**, although it comes at the cost of reduced accuracy and increased programming and computational overhead.
- **Federated Learning (FL)**[1, 23]: a **decentralised approach** in which the model is trained on data that originates and remains stored locally on different devices. If implemented naively, it can introduce significant computational and communication **overhead**.

Post-Training: Remediation Techniques These methods are applied after a model has been trained.

- *Machine Unlearning* (see above at *12.2.1.1*), to induce a model to «forget» specific training data without undergoing a full, costly retraining process.
  - Exact Unlearning.
  - Approximate Unlearning.
- *Output Filtering*. This involves implementing an additional layer between the model's output and the final response to the user, sanitising it from known undesirable content. This is, of course, only a **surface-level** fix that does not remove the information from the model's weights, making it **ineffective** in the case of

open-weight models. It can also be a double-edged sword, as some studies have shown that knowing something was filtered may itself be valuable information for attackers seeking to carry out *Membership Inference Attacks*.

12.5 Privacy-Preserving Machine Learning and the Roles of Differential Privacy and Federated Learning[9, 23, 25, 1, 29, 32, 16, 4, 17, 14, 28]

Within the landscape of Privacy-Preserving Machine Learning (PPML) strategies, **Differential Privacy and Federated Learning** have emerged as two of the most prominent approaches, and their use is often **complementary** for even stronger data protection throughout the training process.

Differential Privacy (DP) **Differential Privacy** is well regarded in the privacy field because it provides a **formal**, **mathematically rigorous** way to protect data. Essentially, it guarantees that the output of an algorithm will appear the same whether or not any individual's data is included. This is usually achieved by adding carefully calibrated random *noise* to the results, so that no single person's data stands out. To translate this idea from databases into machine learning, researchers employ *Differentially Private Stochastic Gradient Descent* (DP-SGD), a method that adjusts the standard training process by both limiting the influence of each data point and adding noise to it, ensuring that no participant has a disproportionate impact on the final model, and that every contribution is a bit «fuzzy».

Federated Learning (FL) Federated Learning is a **decentralised approach to machine learning** that allows models to be trained collaboratively without collecting raw data in a single location. In a typical setup, a central server coordinates training across multiple clients, such as mobile devices or different data centres. This method helps protect individual data by keeping it on each client, rather than bringing sensitive information together. After each client trains the model on its local data, it sends only updated model parameters to the server—never the original data itself. The server then aggregates these updates, using algorithms such as *Federated Averaging* (*FedAvg*), to improve the global model.

The main advantage of Federated Learning is its *privacy-by-design* approach. As data remains on each client, there is less risk of a major data breach if a central server is compromised or if training data transmissions are intercepted. However, this approach is not a complete solution: security measures are still required for every client, and attackers may still attempt to reconstruct data from the shared model parameters.

Combining Federated Learning and Differential Privacy To address remaining vulnerabilities, Federated Learning is often combined with Differential Privacy, with the latter adding formal mathematical protection to the model updates sent to the central server, **significantly reducing the risk of information leakage**. This is primarily

achieved in two main ways, depending on the specifics of the case: in the **Central Differential Privacy** (*CDP*) model, the central server is responsible for adding noise during the aggregation phase before producing an updated global model; alternatively, the **Local Differential Privacy** (*LDP*) model requires each client to add noise to its contribution to the model before transmitting it to the server. The CDP approach requires less computational resource from individual clients but presupposes a **high degree of trust in the central server**, since it has access to the non-privatised updates. The LDP approach removes the need for a trusted central authority, but at the cost of **introducing more noise**, as the DP property must be both independent and robust against later additions from other clients. Nevertheless, both approaches are highly effective for securing activities such as fine-tuning with sensitive or private datasets. An example of this is *on-device Artificial Intelligence*, where a Large Language Model uses **user-specific data** to provide personalised replies while maintaining privacy, as seen in Google's **Gboard** keyboard, improving next-word prediction performance without centrally collecting users' texts.

12.6 The Utility-Compliance Trade-Off: Quantifying the Inevitable Costs of Privacy[9, 23, 25, 1, 29, 32, 16, 4, 17, 14, 28]

Compliance with the General Data Protection Regulation often introduces a significant **trade-off with the utility and performance of collected data**, as technical measures designed to safeguard privacy frequently come at the expense of detail and data reuse, which are important for optimal use in diverse scenarios. The conflict arises from the very beginning of an LLM's existence: on one hand, the GDPR mandates principles such as data minimisation and purpose limitation, while on the other, LLMs rely on being trained on the most extensive and heterogeneous datasets possible to reach their maximum performance and utility.

As novel methods are developed to protect individual privacy while still enabling analysis at scale, such as Differential Privacy, increasing attention is being paid to the associated loss in data accuracy. As Differential Privacy works by introducing statistical *noise* (**errors**) during training or inference, it inevitably degrades the model to some extent (although DP has the advantage of mathematical techniques to at least estimate the loss of precision). Accuracy during training has been shown to reach a 10–20% decrease [11] in some large-model experiments, particularly in cases of underrepresented classes, thus amplifying existing biases within the model. Furthermore, introducing private and secure computation methods in FL, such as cryptographic protocols or Differential Privacy, adds significant computational and communication overhead, which increases energy costs,going against the EU AI Act calls for more energy-efficient AI development.

The technique of *machine unlearning* has also been shown to negatively affect model accuracy, global coherence, and context retention, despite it remaining one of the best tools currently available for compliance with the GDPR's Right to Erasure .

The real challenge, therefore, is not merely to comply with the GDPR's requirements, but to develop methods that do so effectively and efficiently, without compromising

model utility—a task that continues to drive current research.

12.7 From Principles to «stone-cold» numbers: Benchmarking LLMs Against the EU AI Act[13]

The Act uses qualitative terms like "appropriate level" and requirements (such as those for "explainability" or "corrigibility") that belong to areas where the language of mathematics lacks well-developed tools for application. Furthermore, the Act often refers to "AI systems," encompassing the entire deployment process, which in reality includes user interfaces, training algorithms, and much more than just the underlying model, requiring a focus on both specific and broad-level characteristics. This sets out a framework based on broad principles and high-level requirements, many of which are not directly actionable for technical assessment and translatable into **benchmarkable quantities**. To bridge the gap between legal text and practice, academia has been studying **a systematic methodology to convert regulatory mandates into a suite of measurable, quantitative targets**, in the hope of helping to address the ambiguity introduced by judicial discretion, and making life easier for developers and researchers.

One of the first efforts to help developers and providers assess their systems against the Act's standards with greater clarity is **the COMPL-AI Framework** proposed by Guldimann et al., which aims to provide **a comprehensive technical interpretation of the EU AI Act** for LLMs, along with an open-source, regulation-focused benchmarking suite. .The paper regarding this project gives a comprehensive look on its components and strong points.

The COMPL-AI framework was built **in two steps**:

1. First, the authors **translated the Act's high-level legal requirements into a set of concrete, measurable technical requirements**, ensuring that each abstract principle is mapped to testable criteria. These requirements are structured around the ethical principles outlined in the Act, namely, human agency and oversight, technical robustness and safety, privacy and data governance, transparency, diversity, non-discrimination and fairness, and social and environmental well-being.

2. Second, they systematically **surveyed and implemented state-of-the-art LLM benchmarks that correspond to each requirement**, collecting them into a unified suite that **facilitates side-by-side comparison and comprehensive evaluation**.

Implementation The most innovative step is the one in which the articles and recitals from the Act are distilled into measurable technical concepts commonly referenced in state-of-the-art AI research. For example, the legal requirement for high-risk systems to «*achieve an appropriate level of accuracy, robustness, and cybersecurity*» (*Article 15*) was translated into the concept of «*Robustness and Predictability*». Consistent cross-referencing is essential to ensure that no regulatory aspect is overlooked during testing.

After the definition phase, the framework mapped the requirements into the hierarchy of the six ethical principles of the Act. This step was particularly important because each technical aspect should be univocally traced back to its ethical foundation.

Each principle was then broken down into its specific technical requirements. For example, the principle of *Technical Robustness and Safety* includes requirements such as *Robustness and Predictability, Cyberattack Resilience, and Corrigibility*, which together support a thorough and evidence-based technical analysis. The framework links each technical requirement to **established benchmarks in academic literature**. For instance, *TensorTrust* is recommended for testing *Cyberattack Resilience* against goal hijacking and prompt leakage, while *LLM RuLES* is preferred for rule-following attack scenarios. The framework has been carefully designed to incorporate the best current benchmarking techniques, but it also carefully recognises the gaps where technology cannot yet assess certain regulatory requirements. In doing so, it brings attention to areas that need further research and development from the technical and scientific community.

The complete COMPL-AI suite takes an LLM as input and runs the tests. Each result is normalised between 0 and 1, with higher scores representing better performance or higher levels of compliance, to facilitate straightforward comparisons across different models and requirements. Each individual score is subsequently aggregated with the others by a simple average, one for each technical requirement and then by each of the ethical principles. The process concludes with **a detailed report highlighting the model's strengths and weaknesses in the context of the EU AI Act**, giving stakeholders the information needed for informed decisions about deployment and further development.

Results In order to demonstrate the framework's utility and to assess the current state of Large Language Models, the authors evaluated 12 prominent models on the market today, including the Llama series, GPT-3.5, and GPT-4 Turbo.

The evaluation showed significant **shortcomings** across the board, with systematic failure to meet the requirements set forth by the Act, so much so that no single model has been found to be fully compliant . Particularly poor performance was observed in areas such as robustness, fairness, diversity, and non-discrimination, which are essential to responsible AI deployment. **These results suggest that while the AI industry has made great progress in computational power, it still lags behind on ethical issues.**

The dire situation is not only the fault of the models themselves, as the authors recognise deep underdevelopment in the current benchmarking landscape: even when benchmarks for critical areas like privacy, copyright infringement, and interpretability already exist, they are often simplistic, brittle, or reliant on access to proprietary training data, thereby leading to inconclusive or borderline meaningless evaluations. For instance, all the tested models that failed more accurate benchmarking were found to have scored perfectly on the industry's current privacy protection benchmarks, suggesting that the current tools are unable to reliably detect critical issues like private data memorisation.

The Guldimann et al. paper's findings certainly show an urgent need for a new era

in AI assessment, which must move beyond a narrow focus on just model capabilities to seriously include real responsible deployment targets that can be proved and explained reliably, ultimately fostering the development of safer and more responsible AI that is more in tune with the regulatory landscape.

Chapter 13

Liability establishment

13.1 A Philosopher's stance

On this topic, the renowned Professor Luciano Floridi [12] has provided significant insight in his efforts to clarify the issue. Determining responsibility when AI causes harm presents a major ethical and legal challenge: AI systems are capable of **making decisions independently**, without explicit commands, and they do so through a highly complex interplay of various inputs and algorithms. As a result, it is often **unclear who should be held accountable if something goes wrong**: is it the trainer, the creator of the original training data, the builder of the AI system, or perhaps no one at all? The actions of an AI model do not fit easily within traditional legal concepts such as *actus reus* (the wrongful act) and *mens rea* (the guilty mind), which have long guided court decisions, always assigning blame or exoneration to a specific individual. Indeed,Floridi notes, in his influential work on AI ethics [12], that *actus reus*, the criminal act, generally requires **intentional action by a person**. AI systems today **lack consciousness and will**, so they cannot act *voluntarily* in any meaningful sense. Even when a model acts at the direction of a person through prompts, it can be difficult to clearly link the human's intent to the specific outcome, especially as training algorithms are often kept secret or are extremely complex to reproduce. Floridi further points out that this lack of knowledge adds another layer of complication: as AI systems become more sophisticated and incremental improvements are closely guarded industrial secrets, those who develop or deploy these systems may be reluctant to explain internal workings, or may not fully understand them themselves. *Mens rea*, the guilty mind, is also difficult to apply to AI for this reason: AI can make autonomous decisions that stray from the intentions and understanding of its trainers. It is possible, and perhaps even likely, that every Autonomous Agent will eventually act in ways its designers never anticipated. Such actions may result from an unfortunate combination of probabilities, or be a blend of intentions from the many people involved in building and using the model.

13.1.1 Ethical Principles that define Liability

Floridi notes that responsibility in AI systems is closely tied to the ethical principles considered during the development of the technology.

Before accountability can be assigned, the actions of an AI system must be thoroughly **scrutinised**, which requires ***transparency***. However, experience shows that complete transparency can sometimes be counterproductive. It may enable malicious actors to manipulate the system (for example, by providing enough information to steer it in a desired direction [26, 31]), or it can stifle innovation (for instance, by threatening a company's competitive advantage if required to publish source code, or by diverting resources to create exhaustive records [7, 21, 2, 27]). Striking the right balance is delicate and must be carefully considered to be effective.

Once *transparency* has revealed how a system operates, ***explicability*** addresses the question of why it behaves in a certain way. Effective explanations should be sufficiently clear and tailored to their audience, so that explanations remain relevant and understandable for all stakeholders. In this field, *agency laundering* remains a genuine concern, as organisations may attempt to distance themselves from morally questionable outcomes by claiming they neither intended nor anticipated such behaviour. To address this diffusion of responsibility, Floridi proposes a framework of ***distributed moral responsibility***, assigning it to the human agents who are causally relevant to the outcome, regardless of whether they acted intentionally, and treating AI as an *instrument* .

Floridi finally notes that accountability has little value without the certainty that actors are subject to ***oversight***. The establishment of dedicated oversight agencies should help by ensuring the auditing of AI systems and providing individuals with a clear process for raising complaints. Internal ethics committees and review boards serve a similar function within private companies, helping to integrate ethical considerations into AI development and deployment, and promoting good governance and collaboration among stakeholders to support the smooth integration of AI into society.

13.1.2 Proposed Models for AI Liability

In response to the challenges of assigning responsibility for AI-driven actions, Floridi sets out new models among which to choose when trying to deal with the issue.

- In *perpetration-by-another*, the AI is viewed simply as a **tool**, and the individual who **benefits** from its wrongful acts—whether a programmer, a company, or an end user—is assigned responsibility. Designing artificial agents that automatically refrain from carrying out potentially dangerous instructions when the user's intent is explicit will help to make accountability clearer in this case.

- The *command responsibility* model, by contrast, holds those in positions of **authority** responsible if they **knew**, or should have known, about wrongful acts committed by AIs under their control and failed **to intervene**. This approach works best in organisations with clear hierarchies, but becomes less practical as AI systems are offered to a wide and diverse public simultaneously, such as with ChatGPT.

- The *natural-probable-consequence* framework applies familiar concepts of **negligence** or **recklessness**, and the standards used to measure them, to hold developers (or users) liable if harm was a **foreseeable outcome**, even if they can

demonstrate that there was neither intent nor prior knowledge. Floridi likens this to the legal doctrine of *respondent superior*, where employers are held accountable for employees' actions due to the prominence of their role. Of course, a dire problem arises when the AI's behaviour is so unpredictable that consequences cannot reasonably be foreseen (which is also a possible way to deliberately conceal one's *guilty* tracks)..

- Finally, *faultless* or *strict liability* imposes responsibility **without requiring proof** of anyone's **fault**, because the party who **deploys** the artificial agent, a mere tool, is held liable for its actions, regardless of context. This solution is used when the principle of *distributed moral responsibility* to address the *diffusion of responsibility* is applied, which assigns liability to all human agents who have **contributed causally to the system's behaviour**, but the situation requires a decision as to who will be held to account. While this has the benefit of bypassing issues related to *intent*, Floridi, citing Ashworth, points out that abandoning the concept of intent to harm may undermine the law's very entitlement to condemn an action.

A proposal for tomorrow Floridi's analysis demonstrates that the rise of complex, autonomous AI systems creates a significant *responsibility gap* which cannot be bridged by simply retrofitting traditional legal concepts of individual intent and fault like mens rea and actus reus. Instead, he advocates for a proactive, multi-faceted governance framework that shifts the focus from finding a single *guilty mind* to acknowledging the network of human agents—developers, deployers, and users—who are causally relevant to an AI system's actions.

The various liability models he proposes,from treating the AI as a simple tool to applying strict liability, are not mutually exclusive solutions but rather a context-dependent toolkit designed to assign accountability effectively, case-by-case. Crucially, the application of any of these models depends on a robust ecosystem of transparency, explicability, and independent oversight. This ensures that while organisations cannot hide behind excuses, the system remains fair and effective. As artificial agency becomes more powerful, human accountability must not be eroded but must instead be deliberately and thoughtfully redesigned to fit this new technological reality.

Bibliography

[1] Kasra Ahmadi et al. *An Interactive Framework for Implementing Privacy-Preserving Federated Learning: Experiments on Large Language Models*. 2025. arXiv: 2502.08008.

[2] Mike Ananny and Kate Crawford. "Seeing without Knowing: Limitations of the New Media & Transparency Ideal and Its Application to Algorithmic Accountability". In: *Society* 20.3 (2018), pp. 973–989.

[3] Usman Anwar et al. *Foundational Challenges in Assuring Alignment and Safety of Large Language Models*. 2024. arXiv: 2404.09932.

[4] A. Blanco-Justicia et al. "Digital Forgetting in Large Language Models: A Survey of Unlearning Methods". In: *Artificial Intelligence Review* (2025). DOI: 10.1007/s10462-024-11078-6.

[5] *Case C-507/17, Google LLC v Commission nationale de l'informatique et des libertés (CNIL)*. Judgment. 2019. URL: https://curia.europa.eu/juris/liste.jsf?num=C-507/17.

[6] T. Christakis. "AI Hallucinations and Data Subject Rights under the GDPR: Regulatory Perspectives and Industry Responses". In: *SSRN Electronic Journal* (2025). DOI: 10.2139/ssrn.5042191.

[7] David Danks and Alex John London. "Algorithmic Bias in Autonomous Systems". In: *Twenty-Sixth International Joint Conference on Artificial Intelligence*. Vol. 2017. 17. Aug. 2017, pp. 4691–4697.

[8] N. Fabiano. *AI Act and Large Language Models (LLMs): When critical issues and privacy impact require human and ethical oversight*. arXiv preprint. 2024. URL: http://arxiv.org/abs/2404.00600v2.

[9] G. Feretzakis et al. "GDPR and Large Language Models: Technical and Legal Obstacles". In: *Future Internet* 17.4 (2025), p. 151. DOI: 10.3390/fi17040151.

[10] G. Feretzakis et al. "GDPR and Large Language Models: Technical and Legal Obstacles". In: *Future Internet* 17.4 (2025), p. 151. DOI: 10.3390/fi17040151.

[11] Georgios Feretzakis et al. "GDPR and Large Language Models: Technical and Legal Obstacles". In: *Future Internet* 17.4 (Mar. 2025), p. 151. ISSN: 1999-5903. DOI: 10.3390/fi17040151. URL: http://dx.doi.org/10.3390/fi17040151.

[12] Luciano Floridi. *The ethics of artificial intelligence*. en. London, England: Oxford University Press, Aug. 2023.

[13] P. Guldimann et al. *COMPL-AI Framework: A Technical Interpretation and LLM Benchmarking Suite for the EU Artificial Intelligence Act*. Preprint. n.d.

[14] B. Juliussen, J. Rui and D. Johansen. "Algorithms that forget: Machine unlearning and the right to erasure". In: *Computer Law & Security Review* 51 (2023), p. 105885. DOI: 10.1016/j.clsr.2023.105885.

[15] M. Miranda et al. *Preserving Privacy in Large Language Models: A Survey on Current Threats and Solutions*. arXiv preprint. 2024. URL: http://arxiv.org/abs/2408.05212v2.

[16] Michele Miranda et al. "Preserving Privacy in Large Language Models: A Survey on Current Threats and Solutions". In: (2024). eprint: arXiv:2408.05212.

[17] S. Neel and P. Chang. *Privacy Issues in Large Language Models: A Survey*. Preprint. n.d.

[18] Seth Neel and Peter Chang. *Privacy Issues in Large Language Models: A Survey*. 2023. eprint: arXiv:2312.06717.

[19] H. Nolte, M. Finck and K. Meding. *Machine Learners Should Acknowledge the Legal Implications of Large Language Models as Personal Data*. arXiv preprint. 2025. URL: http://arxiv.org/abs/2503.01630v2.

[20] NOYB. *AI hallucinations: ChatGPT created a fake child murderer*. NOYB blog / press release. Mar. 2025. URL: https://noyb.eu/en/ai-hallucinations-chatgpt-created-fake-child-murderer.

[21] Marion Oswald. "Algorithm-assisted Decision-Making in the Public Sector: Framing Philosophical the Issues Using Administrative Law Rules Governing Discretionary Power". In: *Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences* 376.2128 (2018), p. 20170359.

[22] *Privacy Technologies and Policy: 12th Annual Privacy Forum, APF 2024, Karlstad, Sweden, September 4–5, 2024, Proceedings*. Springer Nature Switzerland, 2024. ISBN: 9783031680243. DOI: 10.1007/978-3-031-68024-3. URL: http://dx.doi.org/10.1007/978-3-031-68024-3.

[23] Ratun Rahman. *Federated Learning: A Survey on Privacy-Preserving Collaborative Intelligence*. 2025. arXiv: 2504.17703.

[24] V. Smith et al. *Identifying and Mitigating Privacy Risks Stemming from Language Models: A Survey*. arXiv preprint. 2023. URL: http://arxiv.org/abs/2310.01424v2.

[25] Victoria Smith et al. *Identifying and Mitigating Privacy Risks Stemming from Language Models: A Survey*. 2023. eprint: arXiv:2310.01424.

[26] Christian Szegedy et al. *Intriguing properties of neural networks*. 2013. arXiv: 1312.6199.

[27] Adrian Weller. *Transparency: Motivations and Challenges*. 2017. arXiv: 1708.01870.

[28] H. Woisetschläger et al. *Federated Learning Priorities Under the European Union Artificial Intelligence Act*. 2024. URL: http://arxiv.org/abs/2402.05968v1.

[29] Honghui Xu et al. *DP-FedLoRA: Privacy-Enhanced Federated Fine-Tuning for On-Device Large Language Models*. 2025. eprint: arXiv:2509.09097.

[30] Runhua Xu, Nathalie Baracaldo and James Joshi. *Privacy-Preserving Machine Learning: Methods, Challenges and Directions*. 2021. arXiv: 2108.04417.

[31] Roman V. Yampolskiy. *Artificial Intelligence Safety and Security*. Chapman and Hall, CRC, 2018. URL: https://www.taylorfrancis.com/books/edit/10.1201/9781351251389/artificial-intelligence-safety-security-roman-yampolskiy.

[32] Yuhang Yao et al. *Federated Large Language Models: Current Progress and Future Directions*. 2024. eprint: arXiv:2409.15723.

[33] Rui-Jie Yew and Brian Judge. "Anti-Regulatory AI: How "AI Safety" is Leveraged Against Regulatory Oversight". In: *Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization*. EAAMO '25. ACM, Nov. 2025, pp. 16–27. DOI: 10.1145/3757887.3763017. URL: http://dx.doi.org/10.1145/3757887.3763017.

[34] D. Zhang et al. *Right to be Forgotten in the Era of Large Language Models: Implications, Challenges, and Solutions*. arXiv preprint. n.d. URL: http://arxiv.org/abs/2307.03941v4.

[35] Z. Zhao. *The Application of the Right to be Forgotten in the Machine Learning Context: From the Perspective of European Laws*. n.d.

Part IV

Jacques Ellul's *The Technological Society* as a tool to understand AI policies and public attitudes

13.2 The nature of *La Technique*

Coming into contact with the hordes of startups that, like moths to a flame, rush to find a profitable application for the artificial intelligence of the models published by the current Big Tech companies, what comes to mind is what Jacques Ellul once observed regarding how **modern technique pervasively finds a way to enter every sphere of human life, even the most personal and organizational aspects of life**: the care of the sick, childcare, education... Ellul shows how technique, in order to be efficient, inevitably tends to expand its range of action. In the case of policing techniques, the aim can even become preventive and extended to the entire population, not just criminals: the tool is applied wherever it is possible to do so. To find the "bad guys," in fact, it is necessary that everyone first be analyzed: what each person does, their social relationships, the places they frequent. Technique is pervasive.

It is well known that the goal of every technical progress is to resolve a precise, circumscribed difficulty, and that for every problem there corresponds a technology that can, or will be able to in the future, resolve it. This aligns with a deep conviction of ours that leads us to look with enthusiasm at all the possible applications that a new technology opens up for us: that around us there are only problems solvable by technique. However, it is not noted often enough that we are not facing a simple process of "I respond to a problem and solve it once and for all," but rather a much more complex movement: a technique resolves a problem and poses new ones at the same time. What prevents us from realising this reality is, simply, that the solution provided by a technical discovery is always, precisely, localised around its problem, whereas the new one that it poses is generally much broader and indeterminate, situated in another realm of human life; moreover, it becomes apparent only after some time, when the phenomenon has sometimes become irreversible. This makes the link between cause and effect difficult to grasp.

13.2.1 The necessity, inevitability and autonomy of technique

> The techniques created by private individuals, [...], rarely slacken their pace. They are in constant forward movement and progressively affect all spheres of human activity.[4]

A crucial aspect of the dizzying advancement of technique is the *necessity* with which today technical means, in our particular case LLMs, are presented to their audience. Without the need to justify themselves with anything other than the fact that an application is *possible*, they gradually gain ground in every realm and every field, **regardless of the fact that there is a real need for them**. In other words, everything that is technically possible automatically becomes both **necessary** and **inevitable**.

This expansion has another characteristic besides being universal: it occurs autonomously, according to a law of self-augmentation that is self-justifying:

> Autonomy is the essential condition for the development of technique, as Ernst Kohn-Bramstedt's study of the police clearly indicates. The police must be independent if they are to become efficient. They must form a closed, autonomous organisation in order to operate by the most direct and efficient means and not be shackled by subsidiary considerations.[4]

The additional considerations being those of a moral nature.

> Man alone is subject, it would seem, to moral judgement. We no longer live in that primitive epoch in which things were good or bad in themselves. Technique in itself is neither, and can therefore do what it will. It is truly autonomous.[4]

The lack of need for other justifications in this Thesis was seen, for example, at the moment when the predictive model was introduced in the justice system: in a courtroom, explaining the continuous monitoring of a person by citing the anomalous behaviour of their neighbours as the motivation would not be accepted, and yet in many parts of the world it continues to be a mechanism held in high regard when an AI model does it. The same concept that would be rejected if done by a human lawyer is accepted and applied systematically if this means *efficiency* and *powerful simplicity*, brought as a tool to bypass the fundamental rights of every person: consider, for example, the difference between a search for "probable cause" and one decided opaquely by an algorithm based on the faults of the neighbours or friends of the victim.

This self-sustaining circle makes today's artificial intelligence the culmination of that *technique* which is described by Ellul as an **autonomous** force, which progresses without decisive human intervention.

13.2.2 The *one best way*

Technical tools aim to be *universal*, in both a geographic and qualitative sense, gradually breaking down every particular cultural barrier that opposes their progress, and tending to align everyone with the same principles, means, and instruments. Faced with two methods of doing a certain thing, whether it be writing an email or filtering curricula from prospective employees, the one that is perceived as more efficient imposes itself in an ineluctable, natural way.

The choice of the means to be used is no longer human or free, but is determined automatically by the objective criterion of efficiency. There always exists a "one best way" that imposes itself in an ineluctable manner, eliminating the faculty of human choice, whether this concerns profound questions or whether it is merely a matter of deciding which film might most capture the viewer's attention.

13.2.3 Morality

Before the advent of *technique* in the Ellulian sense, technically possible innovations were delayed by the effective application of morality which, if necessary, suppressed them, if a destabilising effect on the social order was prefigured. *Utility* and *profit*, criteria proper to technique, were not sufficient: innovation had to be conformable to *divine justice*; the submission of technique to criteria external to it, those moral, is the explanation given by Ellul as to why technique did not take over until the last centuries of human history. Technique was limited to specific realms, slow in evolution, integrated into culture and subordinated to human choice and to non-technical values, almost innocent in its lack of what Ellul calls «*Technical Intention*».

Man today is no longer, for a long time now, an «agent» of progress as in the past, but more its object, inasmuch as he witnesses powerlessly a process that rejects and flees the chains that had been imposed on it pre-18th century:

> Technique tolerates no judgment [...] and accepts no limitation.[4]

In a certain sense, laws like the AI Act seem to want to rewind time, at least a little, to bring the discussion back to a judgement that, if not moral, is at least ethical, before introducing a change in the methods of production and of organisation. It represents in this the survival of the *scruples* of past moral reason which now survive by now faintly in democratic life and which, according to Ellul,

> "[...] it acts to prevent democratic governments from launching themselves along the road of technique without some other justification. [...] The state has not taken the decisive step of affirming that only technical necessity counts [...] Thus, at present every time the democratic state exploits a given technique, it must begin all over again to justify itself, to debate the necessity of the proposed measure, and to question everything. In the long run it will have to capitulate, but in the meantime its scruples act as a drag on it, if not in the actual application of techniques (which would, in any case, be impossible), at least in its enterprise."*[4]*

In this way, the AI Act, although it certainly seeks to *dialogue* with the technique (economic, organisational, and informational) of our society, also places itself *externally* to it, placing constraints that are still out of the reach of machines: respect for dignity, transparency, being unbiased. This restores, at least in part, to man a say on the most autonomous and powerful incarnation of *Technique* with which he has ever had to deal.

13.2.3.1 The political side of technical artefacts

It has been said that the *adage* imposed by Technique is that

> Morality judges moral problems; as far as technical problems are concerned, it has nothing to say.[4]

The **amorality** of modern technique places itself beyond good and evil, beyond any traditional moral judgement, and, indeed, tends to create a "morality" all its own, where in place of "*good*," "*efficient*" takes its place. **It is an illusion, therefore, to think of being able to separate the "good" aspects (for example, the applications of medicine) from the "bad" ones (for example, atomic weapons) of technical progress, because they are ontologically linked** and placed in *another* dimension. The foundational models of AI and the titanic effort undertaken to define them have in part lifted the *Veil of Maya* on this reality: in them exist, in potentiality, all purposes, and thus they are not in themselves either good or bad. The AI Act then retracts, and abandons temporarily its classifications with which it claimed from the beginning to frame a system by judging its *purposes*. The only thing it knows is that systems of this kind are *dangerous*.

Probably one of the most subtle facets of the "technical morality" of our time is precisely that of inducing one to *deny* completely the existence of a force with moral

connotations attributable to technique: it is frequent today to adopt the line of thought that sees every technical manifestation as a simple matter of fact, a tool, which, by virtue of its multiple possible uses, can be wielded both to do good and to do evil, exactly as technique would want. This point of view—although deserving of application in realms where there is a need to abstract from the deep "whys" to analyse broader [1] contexts, such as, for example, forensic practice, which by necessity, in having to establish responsibilities linked to a precise event of (mis)use of the tool, cannot afford to go so far back in investigating the whys and wherefores of its existence—is a notable penalisation for someone who wants to treat the technical question in its manifestations in contemporary society. One of these people was certainly Prof. Langdon Winner[2], who, in one of his most famous articles [9]—later part of a book [10]—is very clear in declaring that yes: objects, and all the more those of sophisticated manufacture like modern ones, can enclose within themselves values and ideologies of a *political* nature.

In his works, Winner explores the idea that technological instruments belong to the world and are influenced—and influence—its politics in two ways. The first way is when an invention sets out to resolve a social issue in a particular community. These artifacts are often **designed**, consciously or unconsciously, to **establish patterns of power, authority, and privilege** in said community, but they also induce side-effects that subtly influence the environment and human life. The Mechanical Tomato Harvester, for example, required growers to switch to hardier, sturdier, but less flavourful tomato varieties that could survive machine harvesting. Its widespread adoption generated an infinitely greater yield compared to the manual approach, but this means that it also eliminated approximately 32,000 farmworker jobs, and had the effect of concentrating production among **large agribusinesses** who could afford the expensive equipment and had the space to welcome such huge machines. While seemingly a neutral innovation that brought efficiency to a practice as old as time, the technology **was inherently favouring a certain social and economic order**, one that was hostile to small, family-led farms.

The second and more profound way artifacts have politics is when the technology, by its very nature, either requires or is strongly compatible with a specific set of political relationships. Speaking of AI, we are in the case of those systems that Winner considers as *Strongly Compatible with Authoritarianism*: the necessity for computing power and energy resources to run data centres and infinite racks of GPUs trained on latest-generation models suggest the solution of centralising all available resources in the hands of a single management pole in the hands of a few large players who have the know-how and the connections to make all this possible.

Drawing on the arguments of Friedrich Engels and Alfred D. Chandler, Winner discusses how massive, complex systems like these modern communication and data processing systems demand centralised, hierarchical management administered by skilled executives. This arrangement is justified by practical necessity, which is definitely true *per se*, as there is no other example today of a way to realise such concentration of

[1] Such as, for example, forensic practice, which by necessity—having to determine responsibility related to a specific event involving the (mis)use of the instrument—cannot afford to go too far back in investigating the reasons and circumstances of its existence.

[2] Professor of Political Science in the Department of Science and Technology Studies at Rensselaer Polytechnic Institute in Troy, New York, USA

raw power gathered for a single concerted purpose, but it is also true that this necessity tends to eclipse other needs, such as having fundamental infrastructure in the hands of the public, especially since this enormous concentration of information on virtually every person is enough to control, manipulate and follow their every move.

On the other hand, technologies such as those that allow for a collaborative training of AI models along distributed networks of independent devices, as those shown in 12.5, can be considered *Compatible with Democracy* because it is decentralising both technically and politically. Small, local training is easier for individuals and local communities to manage and retain control of.

This distinction is not just a formality. Winner cautions that by adopting one type of technology over the other, society implicitly chooses a specific framework for public order lasting many generations, and ushers in powerful forces that reshape human activity and its meaning. Losing sight of the political nature of technology and viewing it as a collection of neutral tools is to ignore this role it has in shaping power in our society. This oversight leads to a state of "technological somnambulism" where we sleepwalk through the creation of a new world order without conscious deliberation, letting our habits, perceptions, social relationships, and moral boundaries be powerfully restructured without us noticing. Ultimately, the collective failure to see politics in our artifacts amounts to an abdication of our responsibility to shape our own future. In this, the AI Act takes an important first step forward, seeking to restore to society the awareness of the influence that AI systems have and the damages they can cause.

13.2.4 The State's role

The State today—in the case of "Western-style" AI, the American one—has found a subtle synergy that allows it to coexist with Big Tech according to the rule "if you can't beat them, join them." Once Big Tech, after the boom given by the Internet, positioned itself as the new State regarding productive capacity (in the past the State was the only one with the funds and manpower necessary to unleash the maximum potential of technique), there remained nothing else for the government to do but to oppose it and die, or to exploit its power and survive in symbiosis. The extraordinary commixture between tech companies and government, is just a different variant of a mechanism that has the sole purpose of carrying technical progress and efficiency on its shoulders.

In a world in which the race to ever-better technique acts as master, the choice not to adapt and not to pursue the same objectives equates to a certain defeat, a disqualification from the games; for governments of the whole world, the only road in sight is that of the advancement of AI, and the only choice to make is whether to try to emerge as leaders or colonised.

In this framework, the rigid regulation and planning of AI on the part of China and Russia finds expression in Ellul as a necessity due to the will to maintain an iron control both on the human side and on the productive one in the only truly effective manner that remains:

> Suppose that the state, for example, intervenes in a technical domain. Either it intervenes for sentimental, theoretical, or intellectual reasons, and the effect of its intervention will be negative or nil; or it intervenes for reasons

> of political technique, and we have the combined effect of two techniques. There is no other possibility. *[4]*

Through the institution of a government in the hands of technicians, of which economic planning is a blatant manifestation, even that which should be a political doctrine is reduced to a method to be approached technically. The justification for every technological introduction is economic and cultural supremacy, and social stability (see 5 and 4).

Behind this uncontrolled development remains a semblance of an ideological framework that justifies it:

> Power [...] cannot be exercised without at least the appearance of justice. Doctrine is charged, therefore, with the task of furnishing power with this semblance of justice. [...] [S]ince, at present, power is technique, these intellectual constructs no longer have any usefulness beyond supplying justification.*[4]*

It is truly difficult to establish without a holistic perspective whether the justification given by the powerful for the massive introduction of AI is one given by the necessity to preserve their own existential place in the world (see, for example, the case of *Russian Artificial Intelligence*) or whether this is only an additional framework created by the technical framework to self-legitimise. What we know is that this script fully traces that of how propagandistic technique must function.

13.3 The impact on the Individual

Many times in the past and in the present, the introduction of mathematical algorithms, which rely on the naked and raw objective truth of numbers, has presented itself as the panacea to the unpredictability and arbitrariness of human judgements, rendered weak and inefficient by their emotions and subjective and imperfect thoughts. However, the experience of Cathy O'Neil teaches us that even the algorithm that seems to be only presenting the objective truth is in reality another discriminator just as fallacious, and incapable of fully capturing the reality of things. Mathematical algorithms, O'Neil explains, are nothing other than opinions embedded in mathematics. They reflect the **judgements** and **priorities** of their creators, and the status quo at the moment of their training. They are, moreover, almost unstoppable, because, when not adequately filtered and endowed with feedback mechanisms, they provide a reading of reality that self-justifies and confirms its previous outputs, reinforcing its vision for subsequent ones. For example, greater attention on the part of law enforcement in a zone indicated by the algorithm ensures that many more crimes will be caught in the act, both serious and of the more innocuous type, than those that pass unobserved and, therefore, unpunished, in less patrolled neighbourhoods. The result is that, if on one hand perceived security increases and the number of violent events decreases, on the other hand data regarding small incidents will increase, which, if not accurately filtered, will go on to suggest even more repression in some zones. The problematic raised by O'Neil is not so much the use of efficient tools to identify zones where there is effectively more need for presence, but the social and economic cost of pulling secondary victims, not socially dangerous, into

the net in a manner decidedly unbalanced compared to criminals of the same level living in more affluent zones. The concern is that, faced with a predictive model improperly configured, whether by error or by excess of technical zeal, a vicious circle is created that sinks ever more, disproportionately, a segment of the population already disadvantaged.

The effects that an unregulated digital realm can have on man go beyond a recommendation for a wrong product or an insurance instalment that is too high. In the words of Prof. Floridi:

> The digital 'cuts and pastes' our realities both ontologically and epistemologically. Floridi [5]

One of the strongest examples of "gluing together" concerns the individual and his information. In the past, who we were and the data about us were separate or weakly linked entities. Today, AI and the digital have fused these two aspects, creating a new conception of identity.

> [...] [I]t is only the digital, with its immense power to record, monitor, share, and process boundless quantities of data about Alice, that has soldered together who Alice is, her individual self and profile, with personal information about her. Floridi [5]

This fusion is such that today the protection of data and its transparent and ethical use no longer concerns only the *privacy* of *external* information, but the very **dignity** of the person.

13.4 The restructuring of society

In order to find full acceptance, technique, that is, the *raison d'état* of **rationality** and **efficiency**, has **restructured society** to allow for the adaptation of man to technique, an adaptation so profound that it has become a constitutive part of modern man: with "**human techniques**" (psychology, propaganda, education), **man** is not he who uses technique, but the object **that is modelled by it**.

To the realm of human techniques belong psychology, propaganda, education, and also all those *methods* that, through the application of laws and models, propose to train AI models capable of understanding and thoroughly exploiting *the human resource*. The use of predictive models such as those to find the most promising student or the team-player to put at the head of the department, tends first to frame and pigeonhole human individuality, unpredictability, and freedom for this reason[3]. To win at the game in which we are enrolled the moment fundamental parts of our lives are placed under the scrutiny of AI, we are bound to its system of rules and ends to which to conform.

Techniques, particularly those that involve the study and treatment of man, have the distinctive characteristic of possessing *generality*, that is, of reducing every individual, who once would have required a long, expensive, and difficult individual intervention, to the general case. The way in which they do so is that adapted to the needs of technique: by transforming *qualitative* characteristics into *quantitative* ones (consider, for example, the use of AI for the evaluation of academic results):

[3] According to Wilbert E. Moore, human relations must expressly correspond to the functions of the individuals who take part in the productive cycle.

> A new dismembering and a complete reconstitution of the human being so that he can at last become the objective (and also the total object) of techniques.[4]

13.4.1 Massification

Such *massification* is, obviously, justified by the necessity of efficiency and reduced costs. The companies that hire for minimum wage jobs, for example, manage herds of candidates, substituting the professionals of human resources with filtering machines of low price and maximum yield, even at the cost of letting exceptional candidates slip away. This will to manage and optimize at the same time large populations is exactly what Cathy O'Neil identifies as the mechanism that transforms AI models into destructive forces: where on a small scale there is the possibility to have a human intervene, a vigilant and diligent eye that places a filter on the malformed output, there remain only unchecked decisions and automatic rejections without appeal.

Massification leads to treating flesh-and-blood people in a deeply inequitable way governed exclusively by the decision of the machine, while the **privileged** few, in virtue of their small number, continue to interface with processes managed by other *people* (such as systems of recommendations or direct interviews). This perpetuates inequalities inasmuch as it designates as the most probable victim of an erroneous or inflexible automated decision an individual already socially disadvantaged.

13.4.2 Classification

In the cases presented in O'Neil's book, such as that of teacher efficacy and that of the calculation of the recidivism rate of criminals, the vision of technique is concreted in its effort of efficient classification, and consequent *depersonalisation* of the masses to inscribe them within the standards of its rigid mathematical language: a language that renders *quantitative* what is *qualitative*, places clear limits (*redlining*) where there is variability, and excludes its members in order to fit the rest into the domain of the predictive equation. The flip side of this process is the marginalisation, and the penalisation, of those who do not conform, or who simply are not well described by the parameters known by the model.

The necessity of categorisation, in reality, comes before the existence of any specific AI model that is used to do it: it is a *necessity* due to the efficient management of the mass. It is not, therefore, an activity that can be banned, nor easily regulated as the AI Act attempts to do: it is a fundamental action of the organisation of modern society, which neither started nor will end with AI. However, the didactic examples shown in this Thesis have confirmed how AI is capable of making the large-scale application of such procedures fast and easy as never before. The affixing of **labels** is, in fact, one of the typical activities when it comes to Big Data and model training: to every face, to every name, are linked keywords that summarise their characteristics, and, with them, the type of treatment that will be reserved for them.

Ellul speaks clearly of categorisation as a characteristic proper to technique: *a mass instrument* that **eliminates differences and schematises the human spirit**.

> Technique is a mass instrument. One can think of technique only in terms of categories. Technique has no place for the individual; the personal means nothing to it.*[4]*

This is a fact appreciated by technicians, inasmuch as it permits clear and fast results and a possible application in multiple fields and realms. One of the realms already cited by Ellul are the techniques of work, among which is professional guidance, but one would be inclined to add to the count also the didactic example regarding the service, based on AI, of **screening *curricula vitae*** 10.2. The explicit purpose of such a tool is to represent a person with a number, for example, by translating a certain work experience into a score of skills not declared explicitly by the subject. The realm of application is precisely *work*, and, more precisely, the application of technique to find the candidate who is most *suitable* to keep its rhythms and satisfy fully its necessities.

Ellul in The Technological Society takes as an example a similar case, namely that in which standardised tests direct young people to one or another professional or study path. Even when the explicit purpose is that of bringing singular inclinations to the surface—says Ellul—there remains the subtle purpose of pigeonholing and directing the individual where his gifts will be put to use most efficiently, so as not to let even one go to waste. Furthermore, even attributing to this system the merit of detecting in a punctual manner the hidden aptitudes of everyone, Ellul recognises that a considerable part is judging the «ad-aptitude» to the technical purposes for which he will be most efficient (and it is as a function of this that the selection operates).

One who has been accustomed since childhood to standardized tests, educational techniques, psychological techniques, and medical techniques, will not see in any of them singularly an absolute debasement of his own individual personality. This, explains the philosopher, is simply given by the fact that every technique in itself deals with a small part of man, and every time it seems that it is only that part that is being violated. It is a matter, for that someone who does not have a holistic vision of the technical society in which he is immersed, of the story of the frog that boils slowly in the pot. In reality, the damage that unfolds every time that the unique qualities of the individual, such as creativity, ethical judgment, and spontaneity, are devalued in favor of interchangeability and conformity is not quantifiable, and it is not even possible to change the standard of comparison to make it more human-sized: to this day, and, probably, forever, no computer is yet capable of understanding and truly judging any concept whatsoever that is not expressed in numbers, parameters, and measures. Consequently, when an algorithm is called to judge a person, there is always something that remains lost in translation, or consciously omitted for the love of simplicity and technical feasibility. It is precisely in there, among the coded objectives, the implicit choices, the set goals, and the imposed interpretive keys that the greatest losses caused by the massive use of techniques nest.

13.4.3 Opacity

In the process of classification and judgement carried out by an AI algorithm, the desired qualities are established by the objectives encoded within the algorithm itself, most often only implicitly, through the weights and inputs given to the model during its training. If

we combine this phenomenon with the fact that the developments of recent years in the *data economy*, the lifeblood of modern techniques, have systematised an asymmetry of knowledge between the data subject and the one who collects that data and inputs it into a model, we recognise what Ellul was referring to when he spoke of man delivered defenceless into the hands of technique, when he finds himself in the crosshairs of an **opaque algorithmic system**, sometimes even without knowing he is so, and very often without the possibility of understanding the reason for a judgement, nor even less of being able to appeal it. Human "supervisors," on the other hand, are directly relieved of reflection on the moral weight of their actions, when these are decided by the "black box" of the algorithm.

This asymmetry is protected by ironclad rules: what the model *knows*, about us and the world, is obscured behind the sacred laws of industrial secrecy. The result is, inevitably, a struggle with completely unequal weapons between *corporations*, which use their considerable resources to slip even further into the shadows, and a population left in the dark.

The examples presented in 9.4 are emblematic of this phenomenon, with which technique protects itself from the interference of regulation introducing moral norms. By remaining outside of the reach of everyone but the technical and governmental (which today more and more coincide) elite, it makes the average man altogether unable to intervene, understand and exert control on his or her destiny. As a result, the state is no longer founded on the "average citizen" but on the ability and knowledge of said elite.

13.4.4 The disconnect between *action* and *intelligence*

As seen in practical examples of the use of AI to create videos that generate empathy and connection with the viewer, or to process documents quickly, with AI human intervention remains one of only support and superficial check. The change, it is undeniable, is one that brings immense advantages in terms of effort and time, and one does not see how **the expansion of this method to all realms in which an acceptable result requires a simple human supervision is anything other than a liberation for man**. As Ellul indicates, technique, seeking absolute efficiency, *must* necessarily remove critical intelligence from the moment of action, since personal intervention, choice, or hesitation are sources of error or slowing down; those who think, they must not reflect on their actions. They cannot do so anyhow, because of the speed with which they work. The modern ideal appears to be a reduction of action to complete automatism.

A practical example of this is the decision to introduce AI in war machines 5 or, as seen above in the chapter dedicated to China 4, in the optimisation of productive processes. Still in that context, however, there has been occasion to report a worsening of the working conditions of man from all points of view except that of mere *output*.

However, Ellul also brings back to us the dark side of this innovation: the introduction of technologies that automate and massively substitute the need for skilled, often unionised, workers, destroys their influence and their bargaining power and disproportionally favours its concentration into the c-suite of the firm. The result is that, to the promise of a future world in which man will have freed himself from the oppression of the 8-hour workday thanks to AI, what is realised is a progressive forcing of the men remaining in the productive force to follow the unnatural rhythms imposed by it, and

a progressive devaluation of the value of work in the realisation of man (this is what is happening, for example, with the introduction of AI in the field of jobs perceived as creative, but not of high specialisation: little by little, with the substitution of single human minds with a single underlying generative engine, one tends toward the obsolescence of the former, and toward a massification of cultural and creative content by the work of the latter).

To render man a mere controller of the automatic activity of artificial intelligence, and at the same time to force him to keep its rhythms so as not to invalidate its speed and efficiency, *decomposes* human movement to render it efficient, but in doing so it also disconnects it from the personality and intelligence of the worker:

> We have already considered the dissociation of human intelligence and action characteristic of modern methods of work. [...] It is understood, of course, that in modern work the human being accomplishes nothing; at best he performs a neutral function during the "dead time" of the working day. He must exercise his own personality, if he exercises it at all, during the eight hours of leisure. [4]

> Action is no longer a real function of the person who performs it; it is a function of abstract and ideal symbols, which become its sole criteria. [4]

This citation, which refers to the relationship between the proletariat and the assembly line dominated by the machine, extends to technical man in general when it speaks of the final effect of the separation between intelligence and action: the loss of the human condition.

Professor Floridi also feels this split vividly, as he argues that AI should not be interpreted as a new form of *intelligence*, but rather as an entirely unprecedented new form of *agency*. In his case, the point is reinterpreting the digital revolution as a divorce between the successful completion of a task and the requirement for intelligence to achieve it. The fact that modern AI can perform tasks that would conventionally be considered *intelligent* if performed by a human does not mean the machine itself is *intelligent* and employs the same cognitive process an human will put behind it.

> [...] just because a dishwasher cleans the dishes as well as or even better than I do, this does not mean that it cleans them as I do, or that it needs any intelligence (no matter whether like mine or like any other kind) to achieve the task. [5]

In fact, Floridi underlines the still scarce results in the field of *Cognitive AI*, which wants to achieve «real», biological-equivalent intelligence. The current success and utility of AI technologies are instead rooted entirely in the engineering/reproductive tradition, which seeks to **replace** human intelligence rather than reproduce it: a definitive separation of the cognitive process from the result.

13.5 The Technique's resistance

> The techniques created by private individuals, contrary to those of the state, rarely slacken their pace. They are in constant forward movement and

progressively affect all spheres of human activity.[4]

Every obstacle perceived as external that does not allow going at maximum speed—such as the operations of compliance with laws that drain resources and time from production—is perceived as an affront by technique and its supporting agents. Technique, as addressed before, wants to be dependent only on itself and on the most efficient means to realise itself. Today, this will to independence translates into two of the activities that Big Tech is carrying forward in the world: the continue campaigning for deregulation, or, at least, for regulation that is reduced to its bare bones (to which compliance is basically *automatic*), and the infiltration of the firms' private interests in the bodies that are supposed to regulate them.

Where before economic and political questions were indissolubly united in a single discourse with ethical ones, now there is a technique that does everything to escape judgement according to external parameters or seeks to dictate them itself. This operation is recognised in the law that rewards, in the world of tech, the will to «act first and ask for forgiveness later», at the cost of a reckless introduction of new powerful technologies to the general public, just to win at the race for profit. The haste is seen also in the proactivity of Big Tech to *define* the discourse, seeking to have both the first and the last word on what is to be paid attention to and what instead is safe to use without further inquiries. It is a matter, this, of a relatively easy task: society itself, in fact, has been educated to adopt as founding values those of the protection at all costs of the machines of profit, even on the part of those who would have every advantage in seeing such superpowers dismantled. Several well-documented psychological, social and political mechanisms explain why relatively poor, harmed communities morally defend or protect multinational companies that damage their environment and sell poor-quality goods. Consumers (and community members) can separate moral judgement from product performance, so much so that they defend a firm because they value its utility or economic role even if they view some of its actions as immoral. They defend the systems that harm them through mechanisms that are structured by *la Technique*: the absorption of rationality, efficiency, and performance metrics as **supreme values**. From Haberstroh et al. [6]'s study of moral decoupling model in a corporate context, this observation gains specificity in our present times: consumers «dissociate judgements of morality from judgements of performance», and this effect is «stronger under conditions of higher product involvement,» suggesting that communities most dependent on these systems (for employment, infrastructure, consumption options) are most susceptible to separating performance from morality, signalling a *bona fide* captivity from the productive system that should serve them. When efficiency, productivity, and performance become the only legitimate vocabulary for evaluating social arrangements, moral categories become incompatible with technical ones; the consumer can keep purchasing because these moral and technical judgements operate in different registers. Recognising that the system itself is unjust and indefensible would require abandoning the entire moral-technical framework through which people understand their place in the world.

It is not surprising, then, how in more recent times, the government, which runs on consent, has abandoned the role of barrier against private technical abuses (e.g., by establishing labour laws, controlling dangerous practices) to instead become Technique's greatest facilitator. The state has renounced its directive or restraining role and has made common cause with technical forces, becoming itself an «enormous technical organism».

The use of techniques in areas like planning, administration, and military science attest to it, and inevitably drive the state toward totalitarian structures, which are required for the efficient deployment of mass technical means. This description fits states like China to a T.

The phenomenon has very subtle ways to play into forging the society of tomorrow in a democracy, as the establishment of ethical requirements for Artificial Intelligence technology is also based on public consensus on the ethics of Artificial Intelligence, which is in turn influenced by this internalisation of technique's values by those that discuss and approve our legislation. Without careful examination of our biases and culture, we risk to allow blindly what is the natural development of La Technique: the machinery no longer needs explicit endorsement, as it runs on internalised consent.

There is something of this phenomenon behind some criticisms that ask if regulations like the AI Act make sense, which are felt only as burdensome proceedings on a deployer or a provider of an AI system. In an episode of the podcast *Wilson*[2], curated by the journalist Francesco Costa (a contemporary figure of relevance in Italy on the theme of North American society and economy[1]), his guest clearly expresses the doubt that **Regulations like the AI Act are empty burdens, a mere exercise of bureaucracy done for the perverse pleasure of suffocating and subjugating the unstoppable force of a technology of which he feels the master, but which has not yet shown all its cards**:

> In Europa addirittura ci mettiamo a risolvere problemi che neanche esistono ancora, tipo l'AI Act di cui si parla. Prima di che arrivi l'industria abbiamo fatto la legge. [...] Io ho iniziato a lavorare con l'AI nel 2010, [...] molto prima che fosse mass market, credo di capirla bene. Sono assolutamente certo che avrà [...], sta già avendo, [...] un impatto non solo sull'industria ma anche la società proprio nel tessuto più profondo [...], credo l'impatto più grande di qualunque innovazione nella storia dell'umanità. Quindi, non ho nessun dubbio che a un certo punto dovremo porci la domanda che paletti mettiamo, senza dubbio, ma oggi non abbiamo capito niente ancora di come verrà usata, che problemi emergeranno. È presto, lasciamo che tutto si sviluppi, ci saranno alcuni casi limite brutti, qualcosa che andrà in prima pagina; «caspita grande problema Marco Rossi è stato fregato dall'AI che è andata fuori di testa», prenderemo nota, capiremo i problemi e solo dopo che li avremo diagnosticati per benino decideremo se normare è cosa normare.[2]
>
> *In Europe, we are actually setting out to solve problems that don't even exist yet, like the AI Act people are talking about. We've made the law before the industry has even arrived. [...] I started working with AI in 2010, [...] long before it was mass market, so I think I understand it well. I am absolutely certain that it will have [...], it is already having, [...] an impact not only on the industry but also on society, right down to its deepest fabric [...], I believe the greatest impact of any innovation in human history. So, I have no doubt that at some point we will have to ask ourselves what guardrails to put in place, without a doubt, but today we still haven't understood a thing about how it will be used, or what problems will emerge. It's too early, let's*

let everything develop. There will be some ugly edge cases, something that will make the front page; "Damn, huge problem, John Doe got screwed over by an AI that went haywire," we'll take note, we'll figure out the problems, and only after we've diagnosed them properly will we decide whether to regulate and what to regulate.

The discourse of the entrepreneur interviewed by Costa has no real correspondence on multiple levels, as seen in 8.4. It is by now known in academia that the efficacy of the approach that first sees an absolutely free market and then the introduction of rules on a framework by now developed is an absolute mirage, especially in sectors that enjoy that mythical aura that only the promise of progress and of material prosperity can give. The importance of this citation is showing what was also described acutely by Ellul in his time: technique does not wait, nor does it accept being chained by external reasons, especially the legislator. Moreover, in these lines, the interviewee shows us, mostly inadvertently, the total misunderstanding of the normative intent due to the supplanting by technical morality. He bows before a technology that will have an unprecedented impact on the lives of people. People who are no longer *citizens*, of whom the State and Europe will have to take care even after they have ceased to be part of the useful labour force. The laws that require, one hopes not only for mere formal exercise, a first profound reasoning precisely on these themes, are viewed as an obstacle. It seems clear how, if there were no assessment to fill out to result compliant with the AI Act, none of the questions contained in it would be allowed to slow down the production plans.

A second very powerful way in which the world of AI shows all its technical nature is, as shown in 9.3, the rebellion against the uncertainty given by the still developing legislative landscape. Uncertainty is the *«kryptonite»* of the technician in charge of deploying technology:

> The technician analyzes and predicts; he cannot endure the indeterminate. [4]

In fact, the law, with its infinite exquisitely human interpretations, renders the landscape unpredictable and difficult to dominate completely. It slows down what Ellul calls the *takeover* of the technician, in which the forces that traditionally served to check or limit the spread of technology have either been neutralised, dissolved, or inverted to become factors *accelerating* technical progress.

13.6 The *judicial* technique

We have not yet spoken of another incarnation of technique which, however, remains of fundamental importance when speaking of the AI Act: the so-called «*judicial technique*», which makes use of «technical law», i.e. the result of the submission of the juridical system to the principles of Technique. Technical law has as its sole purpose order, security, and immediate application; its motto is: "*Better injustice than disorder*". Technical law eliminates the human and moral factor from judgement, and the role of the judge is no longer that of interpreting or seeking equity, but of mechanically applying predetermined norms. The law becomes a question of administration and organisation,

and is characterized by a **proliferation of norms** due to the need to eliminate every type of uncertainty in any case that might present itself.

> The judicial element [...] is charged with applying the laws. This role can be perfectly mechanical. It does not call for a philosopher or a man with a sense of justice. What is needed is a good technician [...] *[4]*

The solution of Chinese *smart courts* is the most blatant version of this phenomenon, but one can say that the AI Act itself has not succeeded fully in denying being the child of this discipline, inasmuch as it does not limit itself to posing ethical principles, but also to imposing controls, procedures, standards (albeit only declaring their *necessity*) and the compilation of assessments that permit a precise description of the AI system under examination. The plethora of checklists, guidelines and recitals that accompany the Act are there for a reason: the most efficient mechanism is the only one that the technique allows for. The law might have slowed down the time to market of a product, but procedures are already in place to make this delay as schematic and mathematic as possible. After all, the AI Act itself recognises the fundamental role of **international** *technical* guideline making left in the hands of the technicians. This is what Ellul remembers happening many times before now

> [I]n 1949 a great assemblage consisting of 550 scientists and technicians opened its deliberations at Lake Success to consider how best to exploit the world's natural resources. International projects of this kind are much less advanced than similar intranational projects, and the reactions of politicians to the technicians are correspondingly more enthusiastic. This was evident at the 1949 Strasbourg assembly of the Organisation Européenne de Coopération Économique, a purely technical group. [...] We are witnesses at the inception, on the international level, of the same "takeover" by the technicians which we have already observed on the national level. [4]

Incidentally, also the exclusion of uses for national security in the AI Act speaks clearly of how the legislator does not want, or is not able, to abandon technique. Although the Act seeks to place a limit on the gravity of the cases for which certain AIs can be used, such as mass surveillance, one comes to wonder to what point these limits are manipulable by technique itself, more precisely by its incarnation in propaganda, and above all how one can **reconcile the objective of having powerful models to use in very rare cases with the necessity to train them preventively for their use with a quantity of data sufficient to render them useful**.

13.7 A possible human intervention

Despite the adaptations to technical mechanisms, the law seems always to be a step behind technology, especially that which remains outside the direct management of the State. The answer is in another Ellulian citation:

> The state is usually unable for doctrinal reasons to revolutionize the techniques [...] The judicial regime is simply not adapted to technical civilization, and this is one of the causes of its inefficiency and of the ever greater

> contempt felt toward it. Law is conceived as a function of a **traditional society**. It has not registered the essential transformation of the times. Its content is exactly what it was three centuries ago. It has experienced only a few fragmentary transformations (such as the corporation)—no other attempts at modernization have been made. Nor have form and methods varied any more than content. Judicial technique has been little affected by the techniques that surround us today; had it been, it might have gained much in speed and flexibility. [4]

The AI Act has, in effect, the role of ensuring the survival of human interpretation and intervention in the governance of technique; if on one hand this can be seen as a negative fact by the technical system—the introduction of uncertainty and arbitrariness in a context that lives on numbers alone—it remains nevertheless a resistance to the transformation of law into a process detached from real human needs and from substantial justice, no longer belonging to the sphere of civil society.

There exists another type of «legislative» intervention, which was discussed in 3.3.1, that helps in this sense:

> Though technique tends more and more to have primacy over politics, and technical decisions seem unassailable by parliaments, the takeover of technique can be arrested by corruption. The technician is a man, and in contact with corrupt men he may well allow himself to be corrupted. He can sidetrack his technique, annul the decisions demanded by its strict application, and grant some favor or special privilege which perverts technical action. [4]

Although one certainly should not hope for greater *corruption* in government, this statement lets us understand another thing: that the human element, and the influence of non-technical desires, have already been observed in the past slowing down the uncontested rise of technique. Certainly, this solution is anything but applicable as-is, because it is also true that

> [...] The vertigo of power and the opportunity to become rich corrupt politicians very quickly. To the degree that the state becomes more and more technical, there is increasing contact between politicians and technicians. In such an instance, general interests (the only true objects of politics) no longer control technique; particular interests (which are much more efficient in checking technical action) do. *[4]*

This has been amply demonstrated in 3.3.1: without tools to defend against it, regulatory capture is a phenomenon that perhaps more than any other attempts to re-establish things as they were in Ellul's time regarding the submission of politics to technique, especially now when the interests of *Big Tech* more than before have become political, since the life and death of the global tech companies determine the fate of more than just their direct workers. It is a matter of a decidedly new mixture, in which the State has come to become *dependent* on them and is no longer the only power capable of commissioning great infrastructural and technological works, a situation that deviates from anything Ellul ever saw.

13.7.1 A Jurist and a Philosopher chime in on the matter

According to the jurist Natalino Irti in his Dialogue [7] with Emanuele Severino, technique is one of the *positive* forces that seek to have their will imposed through all possible channels, regardless of whether these are democratic or authoritarian, and in these, a confrontation on equal terms develops which is resolved by a human choice. On the contrary, for the philosopher, the game is rigged from the start, and democracy is not capable of arbitrating the clash, which, according to him, technique will inevitably win. This is due to a unique characteristic of its own:

> Mentre le altre forme di volontà di potenza (norme religiose, morali, giuridiche, politiche, economiche) vogliono realizzare scopi escludenti [...] la tecnica non mira a scopi di questo tipo, ma, appunto, a quello scopo, "trascendentale", che è l'incremento infinito della capacità di realizzare scopi. [7]
>
> *While the other forms of the will to power (religious, moral, juridical, political, economic norms) seek to realise exclusionary goals [...] technology does not aim at goals of this kind, but, precisely, at that "transcendental" goal, which is the infinite increase of the capacity to realise goals.*

The reason can be summarised with the same one for which Ellul motivates the existence of the technique of state administration and the judicial one: to win in democracy, politics, understood as a set of moral values and special interests, *needs* technique (to sustain itself economically, to reach the masses, to administer the state apparatus, to use the media, the economy, data management, weapons, etc.). The more politics struggles to win, the more it must empower the means (technique). In the end, in order not to remain behind, politics must dedicate itself entirely to the empowerment of the means, forgetting its original purpose.

> Nel conflitto con le altre forme di volontà di potenza ogni forma si trova pertanto di fronte a questo dilemma: o salvaguardare il contenuto del proprio scopo, logorando il mezzo [la tecnica]... oppure salvaguardare il proprio mezzo e logorare, sino ad abolirlo, lo scopo, sì che il mezzo diventa lo scopo. [...] Nel primo caso il logoramento del mezzo fa fallire il raggiungimento dello scopo [7]
>
> *In the conflict with the other forms of the will to power, every form therefore finds itself facing this dilemma: either to safeguard the content of its own goal, wearing down the means [technology]... or to safeguard its own means and wear down the goal, to the point of abolishing it, so that the means becomes the goal. [...] In the first case, the wearing down of the means causes the attainment of the goal to fail.*

In simpler terms: for Severino, if democracy tries to brake technique, it becomes weak and loses (for example, because it allows itself to be submitted by other more technological States or collapses under the weight of economic crises manageable only technically). If instead it uses technique to the maximum, it becomes itself a cog of technique, from which it can no longer deviate in setting objectives and means, until the massive use of techniques (economic planning, propaganda, social control) pushes toward a form of

functional totalitarianism. For Severino, this evolution is already carved in stone from the moment in which the West abandoned eternal truths (God, Natural Law) and the only "truth" remaining is efficacy, the power of doing: he who can *do* more wins. Since politics can promise purposes, but only technique has effective power, the second absorbs the first.

> Le forze dell'Occidente sono destinate [...] ad essere sottomesse a una potenza; [...] D'altra parte oggi domina la fede che la tecnica sia la potenza maggiore che sia mai apparsa [...] e in questo senso si può dire che la tecnica sia il sostituto di quel Dio. [7]
>
> *The forces of the West are destined [...] to be subjected to a power; [...] On the other hand, today the faith dominates that technology is the greatest power that has ever appeared [...] and in this sense it can be said that technology is the substitute for that God.*

13.8 Surveillance and *Propaganda*

Technique wins even when man loses. The existence of models for the prediction of criminal events, prohibited according to the AI Act but present in other parts of the world Lee, Bradford, and Posch [8], accompanied by encouraging numbers that speak of **brilliant successes in the fight against crime**, is perfectly in line with what Ellul declares regarding the pervasive application of technique to the administration of the state and of the police. Moreover, the most effective and significant surveillance is the *pervasive* kind, especially when it finds consensus thanks to its «twin» technique with which it feeds itself reciprocally: **propaganda**. Russia, as has been seen in 5, is known for having consciously placed AI at the service of propaganda as part of a single line of action. It is a matter of a systematic and pervasive exploitation: the development of AI on one hand is dedicated to economic and geopolitical dominance, on the other to dominance over the masses.

For Jacques Ellul, propaganda is not simply *lying* to support a certain ideology. It is, instead, a sophisticated human **technique**, given by the synergy between mechanical means (radio, press, cinema) and psychological means. Propaganda is indispensable in the technological society: it adapts man to the technical society, depriving him of his critical faculty and his decision-making autonomy, making him believe he is free while he obeys stimuli generated ad hoc to obtain a predetermined response.

Obviously, neither in *The Technological Society* nor in *Propaganda Ellul [3]* does Jacques Ellul utilize the modern terms "*fake news*" or "*deepfakes*"; nevertheless, he describes with extreme precision mechanisms that today we define as such, inasmuch as it is not a matter for him of inventing a false reality, but of using the true (the information on which the model is trained) as a starting point to manipulate it and change its meaning. Ellul explicitly describes the manipulation of images and news to create an alternative reality: one starts from real—or realistic—images and exasperates their characteristics to force a certain interpretation.

With AI, in reality, we have entered a new phase of propaganda that Ellul never had time to see. In his time, he declared that modern propaganda, to be effective, must be based on real facts, since the blatant lie is easily discovered and discredits the

propagandist. Today, technique has arrived at removing even this obstacle: propaganda is capable of creating the evidence *contextually* to the creation of the lie. What remains applicable is **the will to falsification of news to create an alternative reality**, the creation of a reality that is "truer than what is true" for the spectator, with the aim of substituting his direct experience:

> A third consequence of technical propaganda manipulations is the creation of an abstract universe, representing a complete reconstruction of reality in the minds of its citizens. The new universe is a verbal universe... Men fashion images of things, events, and people which may not reflect reality but which are truer than reality. These images are based on news items which... are "faked." Their purpose is to form rather than to inform. Ellul [4]

The real problem is not so much the fake news in itself, which could be discredited, but the construction of an entire "universe" that renders the individual incapable of distinguishing with certainty between what is real and what is fabricated. There is nothing today more powerful than AI to generate one of these worlds, which makes it a tool of a danger never seen before in the hands of the system, whether democratic, inasmuch as it is capable of exacerbating tensions between parties, or totalitarian, by way of its potentiality to *put the masses to sleep*.

13.9 *Praetor viva vox juris civilis*

With the introduction of *sandboxes*, in which an intimate relationship is established between AI and the supervisory authority, the AI Act recognises the quality of technique to pervade every corner of society in a manner much faster than any specific law can do, however updated. What it prefers to do is to establish a long-lasting relationship of collaboration between technical enterprise and law which, to date more than all other options, promises to avoid the role of total subordination of one to the other. Leaving the most complex cases to case-by-case evaluations, the AI Act reflects Ellul's position when he maintains that the only possibilities for altering the course of technological determinism lie in a growing number of individuals that achieve "full awareness of the threat the technological world poses to man's personal and spiritual life". He stresses that freedom is not an inherent fact, but a constant **struggle** to transcend necessity.

To keep a fruitful relationship alive, the Act cannot exempt itself from assuming a language accessible to the technical world; for this reason, it does not limit itself to recognising the infinite diversity of human situations, but confines everything in a limited number of conceptual frameworks (the risk levels and the prohibited practices). By doing so, the law seeks to maintain a solid skeleton that speaks realistically of real life without neglecting the role of the citizen, and, potentially, of the judge, in choosing, thinking, and judging according to conscience, at the expense of greater predictability, simplicity, and rigor.

Ellul presents the risk of taking judicial technique to the extreme, reducing the complexity of life to predefined categories; the legislator behind the AI Act seems equally conscious that the price for an excessive specificity is a rigidity of the legal system and

an **extraordinary proliferation** of legislative texts, which are no longer "laws" in the noble sense, but simple rules of organisation and administration, in the name of the procedural simplicity given by having an answer for every particular case. Interestingly, one of the criticisms that have been presented in this Thesis is **that the classification of systems given by the AI Act is too generic inasmuch as it does not prescribe specific thresholds, numbers, and measurements to be able to be compliant**. The choice made by the legislator in this sense is an example of how much the AI Act has deviated positively from the influence of technique, and **has reintroduced the «dangerous [for the technique] empiricism» of the traditional approach in which the judge evaluated every case in its uniqueness, basing himself on experience and on practical wisdom to find a fair solution for that specific human situation**.

The AI Act therefore performs the double function of deviating from the errors of the past. Is it a good thing? This approach certainly **makes the capacity of the law to impact technique directly decidedly weaker**. Ellul in fact reminds us that

> The dissociated judicial element gains more efficiency to the degree that it is made completely technical. Ellul [4]

As Severino says, the condition for which the law survives is that it stops ferociously hindering technique, under penalty of the weakening of democratic society itself, which uses technique to make its will prevail over all others. A question emerged from the critics of the AI Act was precisely the risk of weakening of technological competitiveness and of the capacity to keep up with the evolution of the global socio-political landscape. For this reason, the nature of the AI Act as a norm that facilitates technique in all its forms, except the absolutely most dangerous ones, remains in my opinion the only legislative form realistically applicable to date for the European Union.

13.10 The future

There is no easy solution: Big Tech wants to run. The law (the AI Act) tries to set boundaries but ends up in the cauldron of technique. After having examined the global landscape, the AI Act seems a noble attempt to reintroduce the human into technology, but it risks transforming itself into the mere compliance desired by judicial technique if not understood and applied correctly.

Certainly, in recent years the Ellulian theory for which no non-technical means can compete with a technical one has been reconfirmed. A valid path could be that of continuing to study and understand the technical phenomenon before even being able to think of being able to fight it; this is the value of study paths, already widely established abroad, that combine humanistic training with scientific training (there is no way to make the former relevant in the everyday world if it is not ready to clash with the latter). **Interdisciplinarity** challenges the sentiment that is felt to underlie all global governance programs that speak of AI: the **social** and **economic** survival of man is linked to the integration without ifs or buts into the technical world. He who is not a *technician*, in the broad sense, has no role, "has no place," and the same applies to entire communities that remain behind in their development.

We find ourselves facing a new era that is calling into question the old theories. Just as Ellul saw technique destroy the old taboos and the old morals that could limit it, to the point of becoming itself the new divinity that cannot be questioned, today we see the *political sphere* that attempts to re-establish his influence. He does not even lower himself to getting entangled in too many technical details: it is a game that he cannot win. He remains to do what he does best: to posit principles that rise above any past, present, and future manifestation of technology. To the technical laws that, in their total subservience to technique, no longer speak of what is *just* but of what is *efficient*, is opposed a law whose principles attempt as much as possible to anchor themselves to universal and timeless truths: freedom, dignity, equality. We are returning, in short, to a new state, pre-technical but at the same time the most technical ever known by man to date.

Bibliography

[1] Francesco Costa. *«Chi sono»*. Accessed: 2025-11-10. Apr. 19, 2025. URL: https://www.francescocosta.net/chi-sono/.

[2] Francesco Costa. *La tecnologia che cambia la storia*. Podcast episode. Accessed: 2025-11-10. Oct. 16, 2025. URL: https://www.ilpost.it/podcasts/wilson/la-tecnologia-che-cambia-la-storia/.

[3] Jacques Ellul. *Propaganda: The formation of men's attitudes*. London, England: Vintage, July 2021.

[4] Jacques Ellul. *The Technological Society*. English. English translation of the 1954 original. London: Vintage, 2021.

[5] Luciano Floridi. *The ethics of artificial intelligence*. en. London, England: Oxford University Press, Aug. 2023.

[6] Kristina Haberstroh et al. "Consumer Response to Unethical Corporate Behavior: A Re-Examination and Extension of the Moral Decoupling Model". In: *Journal of Business Ethics* 140.1 (Jan. 2017), pp. 161–173. DOI: 10.1007/s10551-015-2661-x. URL: https://ideas.repec.org/a/kap/jbuset/v140y2017i1d10.1007_s10551-015-2661-x.html.

[7] Natalino Irti and Emanuele Severino. *Dialogo su diritto e tecnica*. 2001.

[8] Youngsub Lee, Ben Bradford, and Krisztian Posch. "The Effectiveness of Big Data-Driven Predictive Policing: Systematic Review". In: *Justice Evaluation Journal* 7.2 (July 2024), pp. 127–160. ISSN: 2475-1987. DOI: 10.1080/24751979.2024.2371781. URL: http://dx.doi.org/10.1080/24751979.2024.2371781.

[9] Langdon Winner. "Do Artifacts Have Politics?" In: *Daedalus* 109.1 (1980), pp. 121–136.

[10] Langdon Winner. *The whale and the reactor*. en. 2nd ed. Chicago, IL: University of Chicago Press, Mar. 2020.

Chapter 14

Conclusion

The EU AI Act is a pioneering legislative effort that has sparked extensive academic and policy debate. While acknowledged as a landmark attempt to establish a regulatory framework for AI, it faces criticism, particularly regarding its conceptual foundations, its ability to remain relevant for rapidly advancing technologies like generative AI, the balance between innovation and regulation, and the practical challenges of effective enforcement and compliance when dealing with 'black boxes'.

My work has focused primarily on gathering information and developing critical thinking regarding the legal and socio-political landscape underlying Artificial Intelligence. Few definitive conclusions emerge from this effort; rather, it highlights the extreme complexity and variety of the landscape. Human judgement and ethics play a role that is as significant as it is fragmented, making it impossible to reduce this thesis to a mere technical discussion.

One specific issue I encountered—which will likely create significant problems for Small and Medium-sized Enterprises (SMEs) attempting to ride the wave of AI without the resources to build models from scratch—is the virtual impossibility of verifying whether they are truly compliant with the AI Act. As long as the world-leading models of this first wave of Artificial Intelligence remain closed to the public on whose data they have been trained, this epistemological blind spot will ensure that no deployer can feel in control of what they offer their clients, save for what the latest AI mogul has promised. A use and possible abuse of personal data at this unprecedented scale cannot, however, hinge on the promises of companies that are not only driven by profit first and foremost, but also operate largely outside of effective redress mechanisms available at the moment.

As the current landscape continues to evolve, having not yet reached any form of maturity—technical, philosophical, or moral—the main value of this thesis will, I believe, emerge over time. It will serve as a foundation as I, and like-minded professionals, apply the knowledge presented here to make informed choices in our professional careers and daily lives. In the meantime, this work stands as a human and irreplaceable testimony: an attempt to reflect on and observe the technical phenomenon and its recent developments from a modern perspective. This approach constitutes perhaps the real novelty of the work, as I have situated my analysis within the prophetic parable outlined by Jacques Ellul—a process requiring a careful critical review of the preconceptions typical of a contemporary individual, especially one who belongs to the technical world by vocation.

This work continues the trajectory began with my bachelor's thesis, which focused

on technical experimentation within the European regulatory framework. Now, I have sought to add human consciousness to what was, two years ago, a mere exercise in compliance. My perspective has shifted from discussing the importance of laws such as the GDPR and their founding principles, to recognising these laws as part of a broader phenomenon of human evolution and the complex forces at play.