Gstrein, Oskar Josef; Haleem, Noman; Zwitter, Andrej

**Article**

# General-purpose AI regulation and the European Union AI Act

Internet Policy Review

**Provided in Cooperation with:**
Alexander von Humboldt Institute for Internet and Society (HIIG), Berlin

*Suggested Citation:* Gstrein, Oskar Josef; Haleem, Noman; Zwitter, Andrej (2024) : General-purpose AI regulation and the European Union AI Act, Internet Policy Review, ISSN 2197-6775, Alexander von Humboldt Institute for Internet and Society, Berlin, Vol. 13, Iss. 3, pp. 1-26, https://doi.org/10.14763/2024.3.1790

This Version is available at:
https://hdl.handle.net/10419/300752

INTERNET POLICY REVIEW
Journal on internet regulation

**Volume** 13 **Issue** 3

RESEARCH ARTICLE

OPEN ACCESS

PEER REVIEWED

# General-purpose AI regulation and the European Union AI Act

**Oskar J. Gstrein** *University of Groningen* o.j.gstrein@rug.nl
**Noman Haleem** *University of Groningen*
**Andrej Zwitter** *University of Groningen*

**DOI:** https://doi.org/10.14763/2024.3.1790

**Published:** 1 August 2024
**Received:** 11 December 2023 **Accepted:** 26 March 2024

**Competing Interests:** The author has declared that no competing interests exist that have influenced the text.


**Citation:** Gstrein, O. J. & Haleem, N. & Zwitter, A. (2024). General-purpose AI regulation and the European Union AI Act. *Internet Policy Review*, *13*(3). https://doi.org/10.14763/2024.3.1790

**Keywords:** Artificial intelligence, General-purpose AI, AI Act, European Union, AI governance

**Abstract:** This article provides an initial analysis of the EU AI Act's (AIA) approach to regulating general-purpose artificial intelligence (AI) – such as OpenAI's ChatGPT – and argues that it marks a significant shift from reactive to proactive AI governance. While this may alleviate concerns that regulators are constantly lagging behind technological developments, complex questions remain about the enforceability, democratic legitimacy, and future-proofing of the AIA. We present an interdisciplinary analysis of the relevant technological and legislative developments that ultimately led to the hybrid regulation that the AIA has become: a framework largely focused on product safety and standardisation with some elements related to the protection of fundamental rights. We analyse and discuss the legal requirements and obligations for the development and use of general-purpose AI and present the envisaged enforcement and penalty structure for the (un)lawful use of general-purpose AI in the EU. In conclusion, we argue that the AIA has significant potential to become a global benchmark for governance and regulation in this area of strategic global importance. However, its success hinges on effective enforcement, fruitful intra-European and international cooperation, and the EU's ability to adapt to the rapidly evolving AI landscape.

# Introduction

The popularity of ChatGPT and similar systems based on artificial intelligence (AI) related technologies surprised European Union (EU) legislators in 2023 (Helberger & Diakopoulos, 2023; Zenner, 2023). After OpenAI launched version GPT-3.5 of its general-purpose chatbot based on a large language model (LLM) on November 30, 2022 (Wiggers et al., 2023), the service quickly became hugely popular with an estimated 100 million monthly active users as of January 2023 (K. Hu, 2023). This coincided with increased popularity of similar "general-purpose", "generative", or "foundational" AI systems capable of producing text, code, audio, images, videos, and similar outputs usually based on text prompts of users. These developments have not only raised questions about the disruptive potential of such systems in the working environment – especially when it comes to the creative domain (Eldagsen's award-winning "promptography"; Whiddington, 2023), or functions typically taken up by women (Briggs & Kodnani, 2023; Kundu, 2024; McNeilly, 2023). While EU legislators already discussed the Commission's proposal from April 21, 2021 on the desirable and appropriate regulation of AI (Veale & Zuiderveen Borgesius, 2021; Gstrein, 2022, pp. 756–760), the unprecedented popularity of general-purpose AI sparked a debate on whether it requires specific laws (Hacker et al., 2023; Helberger & Diakopoulos, 2023), or even legally binding international treaties (Harris, 2023; Milmo & Stacey, 2023).

Such requests are not surprising given that general-purpose AI systems are used naively at some times, even maliciously at others, potentially causing significant harm (Blauth et al., 2022). Maham and Küspert propose to divide the risks relating to the use of general-purpose AI in three different categories (Maham & Küspert, 2023, pp. 2–5). According to their report, first there are risks stemming from unreliability and a lack of transparency. This results in discrimination and stereotyping, misinformation, privacy violations (Mukherjee et al., 2023), as well as accidents resulting from the use and reliance on inaccurate information. At the time of writing, for example, there are reports of court cases in the United States (Weiser, 2023) and South Africa (Prior, 2023) in which lawyers used LLMs to produce written submissions containing case law references that were entirely generated by the system and therefore unreliable and irrelevant. Secondly, general-purpose AI systems should be considered as dual-use technology, capable of enhancing cyber-attacks and exacerbating (cyber-)security related risks (Casarosa, 2024; Europol, 2023). They might be used to generate text or code in support of large scale phishing attacks, ransomware attacks, or similar types of cybercrime. However, these systems are also exposed to cyber-attacks themselves. Prompt engineering and prompt in-

jection, as well as other techniques to circumvent ethical and legal safeguards built into AI systems, have proven effective (Perez & Ribeiro, 2022). This opens a whole new domain of risks for AI system providers where non-intended responses of AI systems can be elicited through creative code and prompt injection, potentially circumventing safeguards that AI providers installed to prevent misuse and to fulfil requirements of laws such as the European Union's Artificial Intelligence Act (AIA). Third, there are systemic risks, as the emergence of general-purpose AI is being driven by a few powerful actors with considerable economic resources and expertise, leading to centralisation of power, societal disruption, and ideological homogenisation (Maham & Küspert, 2023, pp. 2–5). A study of 14 LLMs suggests that many popular systems are inherently biased when it comes to the answers they give to sensitive political questions, potentially mainstreaming certain ideologies, as well as fuelling societal unrest and division as their use becomes more widespread (Feng et al., 2023; Heikkilä, 2023).

This article provides an initial analysis of the EU AI Act's approach to regulating general-purpose AI, arguing that it marks a significant shift from reactive to proactive AI governance. While this may alleviate concerns that regulators are constantly lagging behind technological developments, questions remain about the enforceability, democratic legitimacy and future-proofing of the AIA. In this study, we adopt an interdisciplinary perspective by taking engineering, governance, and legal aspects into account. Drawing from these perspectives, Section 1 highlights and summarises relevant technological and legislative developments before delving into definitions of central concepts, such as general-purpose AI, in Section 2. In section 3, we examine the proposed requirements and obligations for the development and use of general-purpose AI. Section 4 focuses on enforcement, governance, and administrative fines related to the unlawful use of general-purpose AI in the EU. In the discussion Section 5, we attempt to reconcile the different disciplinary views presented in light of the compromise found by the legislators. We speculate on how the position of the EU as a proactive regulatory actor in this area affects its influence going forward – along the notion of the “Brussels effect” (Bradford, 2020) – and what potential the adopted stance has to shape the evolving landscape of developing and commercialising the use of general-purpose AI.

Our findings are based on the legislative text adopted by the Council on May 21, 2024 (Council of the European Union, 2024), following the adoption of the final political position on the AIA by the European Parliament on March 13, 2024 (European Parliament, 2024). This final version of the legislative text was published in the Official Journal of the EU on July 12, 2024 as Regulation (EU) 2024/1689, and

will become fully applicable as of August 2, 2026 according to Article 113 of the AIA. The rules relating to "prohibited AI practices" such as social scoring will apply already as of February 2, 2025, and the rules relating to general-purpose AI at the centre of this article apply as of August 2, 2025. To expand our analysis further, we took note of the draft AIA text as adopted by the European Parliament on June 14, 2023 (European Parliament, 2023) – the first draft version of the AIA to fully take into account the significant developments around general-purpose AI. This position also promoted a more fundamental rights centred approach (Chiappetta, 2023, pp. 19–21), which was challenged by several EU Member States (including France, Germany, Italy, and Spain) during the trilogue negotiations held with the parliament and the European Commission in autumn 2023. The approach to the regulation of general-purpose AI was one of the most controversial aspects during these negotiations, which was eventually resolved on December 8, 2023 after an unprecedented 36-hour negotiation marathon resulting in a political compromise (Bertuzzi, 2023b, 2023c). In addition to these institutional developments, our conceptual and doctrinal analysis takes into account the emerging body of testimony from civil society organisations and other relevant stakeholders, as well as the available academic literature, which will be presented throughout the following sections.

## Section 1: Context, framing, and scope

To start with exploring the context of technological developments, Bommasani et al. (2022) classify recent developments in the field of AI into three different categories on the basis of underlying technical approaches employed to artificially mimic intelligence. Starting from the mid 2000s (or even earlier), the first category of AI comprises traditional machine learning (ML) models, which are often trained using an annotated dataset of significant features to learn unknown relationships between input and target variables (Géron, 2019). The learning process determined optimal weights (or values) of related model parameters and in turn enabled prediction type tasks (e.g. classification of unseen data samples or forecasting of unknown values). This AI category is referred to as "task-oriented" or "feature-oriented" models, which often achieved reasonable performance in terms of prediction accuracy when using an optimal quality dataset and substantial model complexity. The limitation, however, is that a new annotated dataset and a new model are required for every unique task, resulting in intensive and time consuming processes (Demrozi et al., 2023). The dependency of the model on the uniqueness of task for which it is trained is often so significant that, for example, a well performing model detecting brain tumours in tomography scans collected from

one medical centre had a significant deterioration in performance when applied to similar types of scans collected from another centre (Garg et al., 2023).

The second regime of AI models started with the introduction of deep learning (DL) in the early 2010s combined with the idea of transfer learning. In transfer learning, contrary to traditional ML models, the DL models were trained on large datasets to perform a certain task but then adapted to perform new tasks through necessary fine tuning of the model parameters, instead of training a whole new model from scratch (Thrun, 1998). Convolutional neural networks were a common type of DL models for computer vision application, which were broadly applied in the form of well known architectures such as AlexNet, VGG, and ResNet to address a variety of tasks with necessary domain adaptation (Alzubaidi et al., 2021). The DL research closely overlapped with the field of computer vision (LeCun et al., 2015), as the models were initially trained on large amounts of imagery data. Later this approach was generalised to other data types referring to text, speech, etc. In addition to adapting to other applications, one prominent feature of DL models compared to first category ML models was their improved performance attributed to (i) the scale of data used for training and (ii) a deeper architecture of the models comprising millions of parameters. The second category of AI therefore remained "model-centric", still focusing on developing better model architectures to outperform its counterparts.

Starting from the 2020s, the third category of AI introduces a "data-centric" approach using foundation models (FM). This category is still developing. Instead of "deeper" models with more powerful architectures, FMs make use of existing ML methods such as supervised, unsupervised, and transfer learning to analyse an unprecedented amount of data through large scale computing. In their own words, the proponents of the FM paradigm proclaim that "transfer learning is what makes foundation models possible, but scale is what makes them powerful" (Bommasani et al., 2022). Technically, two key attributes of foundational models i.e. "emergence" and "homogenisation" distinguish them from previously discussed categories. Emergence refers to the process of achieving a certain behaviour of the model through induction of information within the model rather than explicitly constructing it through, for example, model architecture or model design. This means that the model produces its results as they emerge through knowledge discovery within training data, which subsequently shifts the focus from model design to training data (hence, data centric). Homogenisation refers to the idea that a generic ML model – built through consolidation of methodologies – can be applied for a wide range of applications, instead of developing multiple models for

specific tasks. The consequence of homogenisation is a freedom from model dependency on task, as was the case with the first category of AI. This results in a significant improvement in generalisability through the necessary adaptation to broader applications. Probably the most notable application of such general-purpose systems are LLMs such as OpenAI's ChatGPT, which take user questions as an input prompt and leverage their emergence capabilities to respond to a broad variety of questions in a human-like narrative.

While the powerful potential of next-generation AI has been successfully introduced to the mainstream in 2023, concerns around transparency, security, trustworthiness, fundamental rights (e.g. discrimination and data protection), as well as sustainability and energy use remain open. On March 13, 2024, the European Parliament voted in favour of a Regulation of the European Parliament and of the Council laying down harmonised rules on AI by a substantial majority of 523 votes in favour, 46 against and 49 abstentions (European Parliament, 2024). Following this vote, the text has been checked and edited through the "corrigendum" procedure, before receiving approval of the member states on May 21, 2024 (Council of the European Union, 2024), and being published in the Official Journal on June 13, 2024. This means that the AIA will apply in full from August 2, 2026, with some provisions – 12 months for general-purpose AI, 6 months for unacceptable risk applications (formally "prohibited AI practices"), introduced below – coming into force earlier, as outlined in the introduction (European Parliament, 2024).

Initiatives to regulate or govern AI are also emerging in countries such as the United States, the People's Republic of China, or Brazil (Engler, 2022; Roberts et al., 2023; Schertel Mendes & Kira, 2023). Intergovernmental organisations such as the Organisation for Economic Co-operation and Development (OECD, n.d.), the G7 nations (Schertel Mendes & Kira, 2023), or the Council of Europe – with its "Framework Convention on artificial intelligence and human rights, democracy, and the rule of law" adopted on 17 May 2024 (Committee of Ministers, 2024) – are working on it. Nevertheless, it seems that the EU AI Act with its broad ("horizontal") approach, more detailed provisions, and legally-binding nature takes the most prominent position – at least a position which is not "too modest" (Mökander & Floridi, 2022, p. 508). The AIA seems bound to become a global benchmark for AI regulation. This is certainly also an objective of many EU representatives, as they hope that the AIA will manifest the "Brussels effect" once again, making the bloc the standard-setter in this regulatory domain of strategic importance as was the case with data protection (Gstrein, 2022, pp. 757–758).

When it comes to the scope of the AIA, it is important to emphasise the centrality

of the much-discussed "risk-based approach". The horizontally – in contrast to focusing on specific sectors such as health, insurance, finance, etc. – applicable AIA divides the application scenarios of AI systems into four different categories: unacceptable risk, which leads to a prohibition of the specific AI application; high-risk, which results in increased regulatory requirements and scrutiny and takes up the lion's share of the provisions in the act regulating substantive matters; limited risk applications, which come with certain transparency requirements; and low/minimal risk scenarios, which mean that no specific obligations apply. As the arguments for and against such an approach have already been comprehensively discussed in the literature (Madiega, 2024; Smuha et al., 2021; Veale & Zuiderveen Borgesius, 2021), this article will not delve further into this aspect. Rather, it will focus on the regulation and governance of general-purpose AI, which is not easy to fit in one of these categories since it is by definition not limited to a specific purpose allowing it to attach to a certain risk category.

The original proposal for the AIA was mainly concerned with the harmonisation of the legal framework for the uptake and use of AI systems on the EU single market (the prominent legal basis of the AIA was – and remains – Article 114 TFEU relating to market harmonisation, together with Article 16 TFEU relating to privacy protection). This also means that the AIA focuses on the "placing on the market, the putting into service, and the use of AI systems in the Union" according to Article 1 par. 2 lit. a AIA, which also means that AI systems purely intended for scientific research and development remain outside of its scope (see also recital 25). Nevertheless, this product safety approach drew criticism from civil society and some academics demanding more attention for fundamental rights aspects, specifically potential individual or collective harms resulting from AI deployment (European Digital Rights et al., 2023; Mantelero, 2022, pp. 83–85; Veale & Zuiderveen Borgesius, 2021, p. 112). The European Parliament started an attempt to reframe the objectives of the AIA in this direction throughout 2023. Those who feared that the AIA would break with the EU tradition of emphasising fundamental rights when regulating technology – specifically following the approach of the 2016 EU General Data Protection Regulation (González Fuster, 2014, pp. 213–252) – will welcome this attempt to change the framing. Ultimately, the final text of the AIA can be seen as a hybrid regulation, with most of its provisions still focused on product safety, standardisation, and consumer protection. However, other provisions, such as those relating to unacceptable risks ("prohibited AI practices") in Article 5 of the AIA, venture into the realms of non-discrimination, data protection, and even criminal procedure law when it comes to safeguards against the misuse of "real-time" remote biometric identification systems in public spaces for law enforcement pur-

poses.

## Section 2: Definition of (general-purpose) AI in the AI Act

This article argues that the AIA marks a significant shift from reactive to proactive AI governance, which becomes particularly visible when considering the 68 definitions enshrined in Article 3 AIA. For the better or worse, many of these definitions will become crucial reference points for AI governance going forward. They have been established through a democratically legitimised legislative procedure following a multi-stakeholder consultation process, in contrast to definitions purely focusing on technological developments, or economic opportunities. However, it might not come as a surprise that in absence of a universally agreed scientific definition of AI (Collins et al., 2021, p. 2), and with many of the subdomains of the field such as ML constantly evolving, it remains a complex task to create a useful legal definition of the regulatory subject (Hildebrandt, 2023). Besides the question on how to generally define AI or AI systems, three categories played an essential role in the discussion, namely: FMs, generative AI, and general-purpose AI systems. In a position paper relating to the trilogue negotiations the Future of Life Institute recommended that general-purpose AI systems include foundation models and generative AI systems, which should be reflected in the definitions used as well as the regulatory treatment (Brakel & Uuk, 2023, p. 4).

According to Kai Zenner – the head of Office and Digital Policy Adviser of MEP Axel Voss from the European People's Party (Christian Democrats) – throughout 2023 the Parliament aimed at adopting a "holistic value chain" perspective with the introduction of these new definitions and the accompanying provisions (Zenner, 2023). Through proactive risk identification, enhanced testing and evaluation, as well as increased documentation requirements, the objective is to mitigate the challenges posed by the new AI based systems. In this way it should become possible to address potential harms such as language biases from an early stage of development, regardless of the final application of a general-purpose system. Similarly, increased testing on aspects such as performance, predictability, safety, and cybersecurity should improve reliability. Finally, more available documentation should allow for a better understanding of the functioning of the systems to avoid the "black box society" effect of algorithmic decision-making (Pasquale, 2015, pp. 191–195). The adopted European Parliament's AIA version from June 2023 contains a definition of FM, as well as a definition of a general-purpose AI system. Generative capabilities however are usually being discussed in connection with

FMs and are therefore not a separate legal category. This is also technically valid as generative systems are in fact a subclass of AI systems (Goodfellow et al., 2014). Instead of assigning labels to unseen samples after training (as in case of discriminative models), the generative systems yield the probability of label assignment for a certain instance.

The final AIA text, however, purely focuses on general-purpose AI and no longer makes a distinction between this category and FMs or generative AI. It defines general-purpose AI models in Article 3 par. 63 AIA as displaying "significant generality [...] capable of competently performing a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications". Also, the exception for models used purely for research, development, or prototyping activities is emphasised here. Another definition exists for general-purpose AI systems mentioned in Article 3 par. 66 AIA, which are based on the general-purpose models just defined, yet have "the capability to serve a variety of purposes, both for direct use as well as for integration in other AI systems". This rather general definition will need more interpretation going forward, yet contains a lot of flexibility which might have ultimately been the deciding factor for the legislators. Furthermore, "regular" general-purpose AI systems have to be kept separate from those general-purpose systems using models with "systemic risk". According to Article 3 par. 65 AIA this "means a risk that is specific to the high-impact capabilities of general-purpose AI models, having a significant impact on the Union market due to their reach, or due to actual or reasonably foreseeable negative effects on public health, safety, public security, fundamental rights, or the society as a whole, that can be propagated at scale across the value chain". Hence, the AIA in its final version only refers to general-purpose AI systems and models, which depending on their capabilities and impact are being classified as evoking "systemic risk" or not. This classification is associated with different regulatory requirements, which will be covered in more detail throughout Section 3.

More generally turning to the definition of AI, the original Commission proposal contained an exceptionally broad definition adjustable via a specific regulatory appendix (Veale & Zuiderveen Borgesius, 2021, p. 109). The final text of the AIA now defines AI in the main text without using an appendix. This seems reasonable to avoid having an AI definition which is dependent on annexes, making it significantly easier to understand and more legitimate from a democracy perspective. However, this choice also makes the regulatory text more abstract and less future-proof, since a definition in an appendix can be relatively easily changed by the

Commission in collaboration with technical experts, whereas the change of a definition in the main text requires a cumbersome legislative procedure. Building on the standard established by the OECD in 2023 (Russel et al., 2023), Article 3 par. 1 AIA defines an AI system as a "machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments". This coordination with the OECD could eventually pave the way to easier international alignment and cooperation in AI governance (Bertuzzi, 2023a). Recently, also the Council of Europe in its Framework Convention on artificial intelligence and human rights, democracy, and the rule of law has included a comparable definition in Article 2. The alignment of the OECD with its 38 member states geographically dispersed across Asia, Europe, and the Americas, the 47 Council of Europe member states plus 11 non-member states working together on the AI convention, as well as the EU AIA approach might even signal an emerging international consensus on the definition of AI systems from a policy perspective.

## Section 3: Requirements and obligations

Given the complexity of assigning a specific risk category to general-purpose AI systems due to the versatility of their application, it is perhaps not surprising that a separate chapter, separate from those for unacceptable, high-risk, and limited risk systems, was eventually included in the AIA, ranging from Articles 51 to 56. However, a notable addition to the provisions in this chapter is a transparency requirement in Article 50 par. 2 AIA, which requires "providers"[1] of general-purpose AI systems producing synthetic content (e.g. text, audio, video) to mark it as such in a machine-readable format. This should facilitate automated or human identification of the output of general-purpose AI systems, helping to combat deep fakes and the spread of non-original content. While malicious content creators will likely find ways to circumvent this rule by manipulating the systems they use (e.g. tweaking open-source models, abusing models intended solely for research, or using models deployed outside the geographic scope of the AIA), its existence at least makes it possible to address this not only as an ethical violation, but also as a legal violation. It should be added that Article 85 of the AIA gives any natural or

1. "Provider' means a natural or legal person, public authority, agency, or other body that develops an AI system, or a general-purpose AI model, or that has an AI system or a general-purpose AI model developed, and places it on the market or puts the AI system into service under its own name or trademark, whether for payment or free of charge.

legal person the right to lodge a complaint with the competent national market surveillance authority if they have reason to believe that there has been an infringement of the AIA, such as non-compliance with the transparency requirements for general-purpose AI systems.

Given the reduction in definitions in the final version of the AIA compared to the version proposed by the EU Parliament in June 2023, as outlined in Section 2, it is perhaps not surprising that the provisions on requirements for placing general-purpose AI on the market appear somewhat relaxed. The more fundamental rights focused approach of the Parliament clashed with a more industry friendly approach of the member states during the trilogue negotiations held during the second half of 2023 (Bertuzzi, 2023b). In the final text, Article 53 par. 1 AIA contains four obligations for providers of general-purpose AI systems: first, they have to create and publish a summary about the content used for training of the general-purpose AI model (for this the supervisory authorities should provide a template). Secondly, they have to put in place a policy that allows them to comply with applicable copyright legislation. Thirdly, they have to provide more detailed information and documentation to providers of AI systems who wish to integrate their models in their systems. Fourthly and finally, upon request they need to be able to provide documentation and information to oversight authorities in national countries, as well as the AI office established on EU level at the time of writing. Article 54 AIA contains a comparable set of rules for authorised representatives appointed by providers established in countries outside the EU. These politically agreed high-level requirements have to be more detailed to be practically applicable. On the one hand, the AIA comes with Appendices XI and XII containing more detailed descriptions of the elements that providers of general-purpose AI systems need to report to oversight authorities or other providers who wish to adapt a general-purpose AI system. On the other hand, according to Article 53 par. 4 AIA and Article 56 so-called "codes of practice" have to be developed and periodically adjusted by the newly established AI Office attached to the Commission, which will be covered in more detail in Section 4.

The AIA favours providers of general-purpose AI models that release them under a free and open-source licence. They do not have to provide specific documentation to oversight authorities or providers who integrate their models into their systems, as it is assumed that this information is accessible by default. During the tense political negotiation process taking place throughout autumn 2023, there were rumours that this exemption was also included to favour some European players. For instance, the French start-up Mistral AI is known for providing open-source mod-

els, yet was later criticised for strengthening its cooperation with the American tech giant Microsoft (Hartmann, 2024). Regardless, some measures required now by the AIA are already widely adopted by the respective professional community. For example, open-source providers of popular AI frameworks (e.g. TensorFlow, SciKitLearn, OpenCV) maintain extensive documentation and detailed instruction of use on individual function level (TensorFlow, 2023). The same documentation combined with open availability of computational code is the fundamental requirement for wider collaboration among a large number of developers. Similarly, version management systems combined with code debugging and revisions address quality assurance requirements. This existing practice can be easily translated towards open-source AI models by additionally providing sufficient summary of training data (or metadata) in already existing standardised formats for data management and sharing. Hence, the AIA requirements related to documentation could be viewed as already "ingrained" in the open-source community to a large extent. Similarly, commercial providers of AI frameworks (e.g. SAS Analytics Software & Solutions) also provide comprehensive user documentation, often alongside interactive multimedia training for user base to provide sufficient instruction of use (SAS, n.d.). Given the commercial nature of their products, dedicated teams of professionals maintain such systems through necessary updates and releases.

Nevertheless, both open-source and regular general-purpose AI models have to comply with additional requirements in cases where their models are being classified as resulting in "systemic risk". The criteria for this classification remain both vague and complex, and will need to be interpreted and updated by regulators along the 7 criteria provided in Annex XIII. Article 51 par. 1 AIA only loosely mentions concepts such as high-impact capabilities based on an evaluation of tools and methodologies including indicators and benchmarks, whereas par. 2 in contrast seems overly precise by stating that any model trained by means of a cumulative amount of computation power – greater than 10^25 floating points precisely – should fall into the systemic risk category. When considering this regulatory approach altogether, one can hardly avoid the impression that the legislators at some point stopped to consider technical details, and just put in the AIA what they had to in order to give executive authorities the means to start investigations relating to the biggest and most capable models currently existing, and in addition to demand from the providers of those models that they take extra care when putting them on the market. This shift in power and responsibility becomes even more visible as, according to Article 52 par. 1 AIA, the providers of those general-purpose AI models with systemic risk themselves need to notify the Commission that they meet the criteria. Alternatively, the Commission may decide to designate

a model as presenting systemic risk.

While all of this may sound dramatic when considering legal certainty, closer scrutiny of Appendix XIII implies that the legislators try to target the very big and influential actors with these provisions, e.g. those having at least 10.000 registered business users, or those with access to very large amounts of computing power. Additional transparency on which models are considered systemic risk should be provided through a list that has to be published by the Commission in accordance with Article 52 par. 6 AIA. In conclusion, it remains to be seen how this dynamic set out by the AIA unfolds between the business community and the oversight bodies, namely the AI Office, the Commission, and the national oversight authorities. A lot of details still need to be clarified, and the effectiveness of the approach will ultimately depend on actual enforcement practice.

According to Article 55 AIA providers of general-purpose AI models with systemic risk have increased obligations to perform model evaluations and document adversarial testing, need to assess and mitigate potential risks such as bias and discrimination, need to document and report to the AI Office or national oversight authorities whether serious incidents took place, and ensure an adequate level of cybersecurity. All of this requires constant assessment and consideration of the potential risks that might be associated with the commercial use of AI models . Advancements in prompt engineering and prompt injection pose significant challenges to complying with these obligations, particularly when it comes to ensuring safety and cybersecurity. Prompt injection, a vulnerability in LLMs, allows attackers to use malicious prompts which force a model to ignore its original instructions or perform unintended actions and leads to unauthorised access, data breaches, or manipulation of the model's responses (Shah, 2023). For instance, a data leak incident relating to Samsung highlights this issue. Samsung employees shared confidential data with OpenAI's ChatGPT, leading to sensitive information like semiconductor equipment measurement data and source code becoming part of the AI's learning database, eventually accessible to anyone using ChatGPT (Petkauskas, 2023). This incident underscores the potential for data breaches when using LLMs to review sensitive information, and the increasing obligation of providers of such models to make users aware of potential consequences in light of AIA adoption.

## Section 4: Enforcement, governance, and fines

The product standardisation approach at the core of the original Commission proposal for the AIA raised concerns around enforcement and finding the appropriate balance between a centralised/decentralised governance approach, as well as the

reliance on either public or private (e.g. self-certification) actors guaranteeing compliance (Veale & Zuiderveen Borgesius, 2021, pp. 111–112). In the final text, it seems that the legislators have opted for a model that centralises coordination, multi-stakeholder exchange, and expertise close to the Commission. Concretely, the facilitation of the implementation of the provisions of the AIA will have to be done by the newly established AI Office, which is attached to the Commission. The setup process of the AI Office started at the end of January 2024 (Kroet, 2024), which seems necessary in light of the fact that some provisions of the AIA start to become legally binding 6 or 12 months after conclusion of the legislative process and publication in the Official Journal. At the time of writing, there are concerns about whether enough experts can be found for the AI Office and whether the new structure will be able to work effectively with such time constraints (Gkritsi, 2024). Furthermore, at the Union level the European Data Protection Supervisor will play an influential role overseeing agencies such as Europol, Eurojust, and Frontex (see e.g. Art. 70 par. 9 AIA). Next to the roles for the new AI Office and the European Data Protection Supervisor, an Advisory forum representing multiple stakeholders including industry and civil society will be established (Article 67 AIA), as well as a scientific panel of independent experts (Article 68 AIA).

Most enforcement powers – with the notable exception of the enforcement of provisions relating to requirements and obligations concerning general-purpose AI as set out in Articles 88-94 of the AIA – will remain with member states and will be coordinated at national level through designated market surveillance authorities. The enforcement structure between member states will most likely differ strongly and become quite complex overall, as it is possible to have multiple national competent authorities. Only the designation of a single point of contact is necessary (see Article 70 par.1, 2). Hence, within member states and according to their different national administrative traditions, it is possible that the governance and enforcement will be divided across several existing or newly established authorities, as long as there is a single point of contact. It remains to be seen how data protection authorities try to position themselves in this debate. For instance, the Netherlands have already made clear that their data protection authority will also be influential in governing and enforcing the AIA (Autoriteit Persoonsgegevens, 2023), and the German data protection authorities have started to argue for a prominent role going forward (Krempl, 2024). The developments in the different member states should then be coordinated on European level through the newly established European Artificial Intelligence Board (Article 65 AIA), which additionally has the tasks to collect and share technical expertise and best practices, develop advice on the implementation of the AIA, as well as issue recommendations and

written opinions, among others (Article 66 AIA).

Reflecting on the proposed enforcement structure for a moment, and borrowing from the mixed experience in putting into place a comparable system for the enforcement of the 2016 European General Data Protection Regulation, critics might fear that too much competence remains in the nationally fractured domain of member states, which might have different political and economic priorities, different levels of AI related skills and literacy, as well as administrative traditions. To give a concrete example, over the last years the Irish data protection authority has been heavily criticised for being too cautious when engaging with American big tech companies, many of which have their European subsidiaries based in Ireland, clearly benefiting the local economy. This led to significant conflicts with the other national data protection authorities (e.g. relating to Meta and its policies around user consent for data collection), which were eventually moderated and settled through a dedicated procedure of the central European Data Protection Board (Daigle & Khan, 2020, pp. 20–21; Li & Newman, 2022, pp. 1707–1714; European Data Protection Board, 2023). Since the currently proposed enforcement structure for the AIA shows a similar decentralised pattern, and given that countries such as the Netherlands already indicated that their national data protection authority will be responsible for the supervision of algorithms and the AIA (van der Beek, 2023), it might not be unreasonable to be concerned that similar national differences in the interpretation and enforcement of the AIA will take place in the future. It is therefore notable that when it comes to enforcing the provisions relating to general-purpose AI, the legislator has placed the emphasis on more centralised enforcement around the newly established AI Office and the Commission, perhaps on the assumption that very specialised knowledge and skills are required to carry out these investigations and procedures.

Finally, to briefly address the potential fines for infringement of AIA provisions (Article 99), not respecting the rules around unacceptable risk can result in fines of up to 7% of the worldwide annual turnover for the preceding financial year for undertakings, or alternatively maximum 35 million Euros of an administrative fine. This should probably be regarded as the consequence for unacceptable practices such as the creation of biometric databases by scraping the web as done by Clearview AI (Dul, 2022), or nudging and manipulating voters along the lines of the Cambridge Analytica scandal (M. Hu, 2020). There are also types of fines which are lower and address less severe infringements of the AIA. Most relevant here is Article 101 AIA, which addresses fines for providers of general-purpose AI models and where not the member states but the AI Office in collaboration with the Commis-

sion plays a central role in model evaluation, especially when it comes to models categorised as systemic risk. In cases where the providers of those models infringe provisions of the AIA, supply incorrect, incomplete, or misleading information, or fail to cooperate with the AI Office and the Commission during an evaluation of the model potentially resulting in corrective measures or limiting market access, the Commission may, after a hearing, impose fines not exceeding 3% of annual total worldwide turnover in the preceding financial year, or a fine of 15 million Euros – whichever is higher. However, it should be noted that this Article 101 AIA will only apply as of August 2, 2026, whereas the other rules relating to general-purpose AI will apply 12 months earlier as mentioned in the previous sections (see Article 113 lit. b AIA at the end).

## Section 5: Discussion: From reactive to proactive AI governance?

AI Governance is probably one of the areas that most clearly demonstrate how challenging it has become to create respected and legitimate regulation in an increasingly complex and diverse society, where perceptions of respect, fairness, sustainability, and justice constantly shift (Taylor, 2023). Not only is the technology constantly changing and evolving in its capabilities but also the expectations about what it is supposed to achieve and for whom. In this context, it is worth carefully reflecting on the applicability of the AIA. Most notably, its rules will not apply to the area of national security and in particular the military sector (Bertuzzi, 2023c), despite this sector being historically one of the main drivers in its development (Morgan et al., 2020, pp. 3–5). Even more controversial than the regulation of general-purpose AI during the trilogue negotiations was the regulation of AI in the law enforcement context (Zellinger, 2023), an area which currently lacks guidance as experiments relating to facial recognition and similar applications, such as emotion sensing, proliferate in many EU member states, such as France (Jasserand, 2023).

Despite these exceptions, many companies will find the stringent and horizontally applicable rules of the AIA limiting (Abecasis et al., 2024, pp. 24–25). Does the adoption of the AIA mean that the EU legislators stifle innovation, thereby hindering economic and societal progress? Given the complexity of the multi-dimensional AI governance landscape, this question fails to address the real issue at hand. Instead, one could ask whether innovation in a datafied society should only be driven by what is technically possible, leveraging short-sighted industry and economic objectives. In particular, general-purpose AI needs a more comprehensive vision.

In this sense, the adoption of the AIA could be seen as a shift from reactive to proactive AI governance, where democratically legitimised regulators – based on multi-stakeholder input – deliberate about and adopt definitions, establishing binding principles as guard-rails to steer AI market adoption. Here it should be reiterated that the AIA largely exempts general-purpose AI related research from its scope. In other words, the impact of AI and in particular general-purpose AI technologies on society has become too important to only have it steered by a few powerful players when it comes to reaping the economic benefits, and the response of the EU is to leverage its internal market regulation power to create legally binding rules which go beyond ethical principles.

From an EU-internal perspective and with the votes for the European Parliament having taken place in June 2024, regulation of AI systems seemed rather uncontroversial with a view to enhancing safety and trustworthiness. From an EU-external perspective, the already mentioned legislative efforts in the United States or the People's Republic of China, as well as events such as the 2023 AI Safety Summit organised by the government of the United Kingdom on November 1 and 2, 2023 (Sparkes, 2023), make it necessary for the bloc to strategically position itself in the global landscape. Considering these two perspectives, the shift from reactive to proactive AI governance seems to be an obvious choice, as it is quite likely that the EU legislators would also face severe criticism if they were simply not responding to the developments around general-purpose AI and leaving the definition of this space entirely to industry and big tech companies.

However, the adoption of the AI Act and the rules for general-purpose AI also comes with many complex questions which need to be addressed in the years to come. First, the enforcement and enforceability of the rules requires the development of sound administrative and market surveillance practices. This means that on the Union level new bodies such as the AI Office need to be adequately – both in quantity and quality of officials working there – staffed and integrated. Authorities at member state level need to be defined and governance mechanisms modelled to provide the right expertise to address complex standardisation and certification issues and to provide easily identifiable points of contact and exchange for system providers. Next to that, the different efforts in the member states and at the Commission level need to be coordinated by the newly established European Artificial Intelligence Board, which also needs to develop guidance in the interpretation of the rules and many Appendices of the AIA. However good and forward-looking some of the substantive provisions of the AIA may be considered by the legislators who adopted it, without effective enforcement and governance, the

adoption of the AIA could ultimately lead to frustration among a population that sees the Union failing to deliver on its promises.

Secondly, while the AIA may have considerable democratic legitimacy and authority mainly stemming from the broad consultation processes that have taken place at the initial development stage (Gstrein, 2022, pp. 756–758), it remains to be seen whether this will continue to be the case. Already the trilogue negotiation process in autumn 2023 was enormously stressful and required the adoption of many provisions and the making of considerable compromises in little time, as it seemed possible that the AIA would never be adopted (Bertuzzi, 2023b). On top of that, many of the current rules in the AIA – such as those relating to general-purpose AI and the systemic risk category – require detailed interpretation by the European Commission and the AI Office. Such delegation of powers comes with challenges to democratic legitimacy, as the actual decisions will eventually be taken by technocrats and not elected representatives. To name just one example to illustrate the challenge that comes with this approach, the phase-out process of the incandescent light bulb technology between 2009 and 2013 was based on broad guidelines set in the so-called Eco-Design directive adopted by the legislators and further interpreted and implemented by the EU Commission. Once regular citizens realised they needed to get rid of their seemingly beloved (conventional) lightbulbs – because "the Commission" decided this – and replace them with more energy-efficient versions, this sparked a significant backlash policy-makers had to deal with (Stegmaier et al., 2021, pp. 16–19). The legislators might have learned from this experience since Article 97 AIA contains reporting duties for the Commission, a potential sunset period for power delegation of five years, as well as the possibility of revoking the powers from the Commission for both the Parliament and the Council.

Thirdly, and finally, probably the biggest uncertainty relating to the adoption of the AIA and its rules addressing general-purpose AI relate to the future-proofing of the act. It remains unforeseeable how the field will develop and to which extent the EU as a regulator will be able to influence this development. The future-proofing of the AI Act needs to focus specifically on general-purpose AI and foundation models, as these types of AI are the most likely to be covered by the AIA exception for military and national security applications. This dual focus provides both advantages and disadvantages for governance. With proper guidance from the European Commission and the AI Office, it is possible to ensure that general-purpose AI and foundation models are developed with built-in ethics-by-design principles. However, the scope of regulatory interpretation by the Commission and the AI Of-

fice may be constrained, as military and national security applications of these technologies could be siloed off into parallel processes. This bifurcation necessitates clear delineation and robust oversight to balance innovation with ethical considerations and security requirements.

Certainly, the regulators will hope that, similarly to the Brussels effect manifesting in data protection law with the adoption of the 2016 EU General Data Protection Regulation (Bradford, 2020, pp. 131–169), the AIA might become a regulatory "gold-standard" which will be copied in other jurisdictions around the world. However, data protection and AI regulation are clearly two very different fields, as the origin of the former lies in the 1970s, having had the time to develop incrementally over decades, also through an EU directive adopted in 1995 (Gstrein & Zwitter, 2021, pp. 2–3). In other words, the principles and individual rights at the core of data protection law were already well established once the work on the General Data Protection Regulation began. When it comes to AI and in particular general-purpose AI, the field will most likely continue to develop very dynamically, and the position of Europe in comparison to other powerful actors such as the United States or China seems to be clearly less influential. It therefore probably does not come as a surprise that the biggest concern of many European corporations is not per se regulation in the form of the AIA but the lack of accessibility of capital to develop products and services (Abecasis et al., 2024, p. 25). Recently this was also confirmed by a report of the European Court of Auditors assessing the EU strategies around research and investment in AI to become a leader in the field. It concluded that the measures by the European Commission and the member states were not effectively coordinated, and that the investment in AI did not keep pace with their global counterparts (European Court of Auditors, 2024, pp. 4–6).

## Conclusion

The EU AIA represents a paradigm shift in the governance of AI, moving from a reactive to a proactive regulatory framework. This shift aims to address the rapidly evolving capabilities of general-purpose AI and ensure that regulation keeps pace with technological advances. As a hybrid regulation, the AIA combines many different areas of law, with a predominant focus on product safety and standardisation, although elements of fundamental rights protection and even criminal procedural law can be found within this broad framework. The European Parliament's attempts to put fundamental rights at the core of AI governance ultimately had to be compromised with the Commission's initial attempt to stick to a consumer protection focused legislative tradition which finds its legitimacy in internal market poli-

cy, as well as the more industry and research-friendly perspective of the member states. Otherwise it might not have been possible to present the final act before the end of the legislative period. In this sense, it could be argued that the AIA seeks to balance innovation with ethical considerations. Nevertheless, complex challenges remain in terms of enforceability, democratic legitimacy, and future-proofing.

The implementation of the AI Act will face many challenges, particularly in terms of enforcement and governance. The establishment of the AI Office, together with the European Artificial Intelligence Board, is essential to provide the necessary expertise and coordination between member states. Effective enforcement will depend on the ability to harmonise the different regulatory landscapes across the EU and ensure consistent application of the rules. In addition, the future-proofing of the AIA will require continuous adaptation to technological advances and emerging risks. Establishing the AIA in the absence of a universally accepted scientific definition of AI will remain a significant challenge, especially given the evolving nature of sub-areas such as ML. These efforts, in line with international standards such as those of the OECD, highlight the complex interplay between technical understanding and legal categorisation. In addition, the exclusion of military and national security applications from the scope of the AIA requires clear delineation and robust oversight to avoid regulatory loopholes. The EU's proactive stance, if successfully maintained, could set a global benchmark for AI regulation, influencing practices beyond its borders and promoting a balanced approach to AI governance. Looking ahead, the AIA has significant potential to set a global benchmark for AI regulation. However, its success will ultimately depend on effective enforcement, fruitful intra-European and international cooperation, and the EU's ability to adapt to the rapidly evolving AI landscape.

## References

Abecasis, P., De Michiel, F., Basalisco, B., Haanperä, T., & Iskandar, J. (2024). *Generative artificial intelligence: The competitive landscape* (pp. 1–35) [White paper]. Copenhagen Economics. https://copenhageneconomics.com/publication/generative-artificial-intelligence-competition/

Alzubaidi, L., Zhang, J., Humaidi, A. J., Al-Dujaili, A., Duan, Y., Al-Shamma, O., Santamaría, J., Fadhel, M. A., Al-Amidie, M., & Farhan, L. (2021). Review of deep learning: Concepts, CNN architectures, challenges, applications, future directions. *Journal of Big Data*, *8*(1), 53. https://doi.org/10.1186/s40537-021-00444-8

Autoriteit Persoonsgegevens. (2023). *AI & algorithmic risks report Netherlands* (Report Winter 2023/2024; pp. 1–45). https://www.autoriteitpersoonsgegevens.nl/uploads/2024-01/AI%20%26%20Algor

ithmic%20Risks%20Report%20Netherlands%20-%20winter%202023%202024.pdf

Bertuzzi, L. (July 3, 2023a). AI Act: Spanish presidency sets out options on key topics of negotiation. *Euractiv*. https://www.euractiv.com/section/artificial-intelligence/news/ai-act-spanish-presidency-sets-out-options-on-key-topics-of-negotiation/

Bertuzzi, L. (November 29, 2023b). AI Act: Spanish presidency makes last mediation attempt on foundation models. *Euractiv*. https://www.euractiv.com/section/artificial-intelligence/news/ai-act-spanish-presidency-makes-last-mediation-attempt-on-foundation-models/

Bertuzzi, L. (December 9, 2023c). European Union squares the circle on the world's first AI rulebook. *Euractiv*. https://www.euractiv.com/section/artificial-intelligence/news/european-union-squares-the-circle-on-the-worlds-first-ai-rulebook/

Blauth, T. F., Gstrein, O. J., & Zwitter, A. (2022). Artificial intelligence crime: An overview of malicious use and abuse of AI. *IEEE Access*, *10*, 77110–77122. https://doi.org/10.1109/ACCESS.2022.3191790

Bommasani, R., Hudson, D. A., Adeli, E., Altman, R., Arora, S., von Arx, S., Bernstein, M. S., Bohg, J., Bosselut, A., Brunskill, E., Brynjolfsson, E., Buch, S., Card, D., Castellon, R., Chatterji, N., Chen, A., Creel, K., Davis, J. Q., Demszky, D., … Liang, P. (2022). *On the opportunities and risks of foundation models* (arXiv:2108.07258). arXiv. https://doi.org/10.48550/arXiv.2108.07258

Bradford, A. (2020). *The Brussels Effect: How the European Union Rules the World*. Oxford University Press. https://doi.org/10.1093/oso/9780190088583.001.0001

Brakel, M., & Uuk, R. (2023). *AI Act trilogue* (pp. 1–13). Future of Life Institute. https://futureoflife.org/wp-content/uploads/2023/07/FLI_AI_Act_Trilogues-1.pdf

Briggs, J., & Kodnani, D. (2023). *Global economics analyst: The potentially large effects of artificial intelligence on economic growth* (Economics Research, pp. 1–20) [Report]. Goldman Sachs. https://www.gspublishing.com/content/research/en/reports/2023/03/27/d64e052b-0f6e-45d7-967b-d7be35fabd16.html

Casarosa, F. (2024). The risk of unreliable standards: Cybersecurity and the Artificial Intelligence Act. *Internet Policy Review*. https://policyreview.info/articles/news/cybersecurity-and-artificial-intelligence-act/1742

Chiappetta, A. (2023). Navigating the AI frontier: European parliamentary insights on bias and regulation, preceding the AI Act. *Internet Policy Review*, *12*(4). https://doi.org/10.14763/2023.4.1733

Collins, C., Dennehy, D., Conboy, K., & Mikalef, P. (2021). Artificial intelligence in information systems research: A systematic literature review and research agenda. *International Journal of Information Management*, *60*. https://doi.org/10.1016/j.ijinfomgt.2021.102383

Committee of Ministers. (2024). *Council of Europe adopts first international treaty on artificial intelligence* [Press release]. Council of Europe. https://www.coe.int/en/web/portal/-/council-of-europe-adopts-first-international-treaty-on-artificial-intelligence

Council of the European Union. (2024). *Artificial intelligence (AI) act: Council gives final green light to the first worldwide rules on AI* [Press release]. https://www.consilium.europa.eu/en/press/press-releases/2024/05/21/artificial-intelligence-ai-act-council-gives-final-green-light-to-the-first-worldwide-rules-on-ai/

Daigle, B., & Khan, M. (2020). The EU General Data Protection Regulation: An analysis of enforcement trends by EU data protection authorities. *Journal of International Commerce and*

*Economics*. https://www.usitc.gov/staff_publications/jice/eu_general_data_protection_regulation_analysis

Demrozi, F., Turetta, C., Al Machot, F., Pravadelli, G., & Kindt, P. H. (2023). *A comprehensive review of automated data annotation techniques in human activity recognition* (arXiv.2307.05988). arXiv. https://doi.org/10.48550/arXiv.2307.05988

Dul, C. (2022). Facial recognition technology vs privacy: The case of Clearview AI. *Queen Mary Law Journal*, *3*, 1–24. https://qmro.qmul.ac.uk/xmlui/handle/123456789/80559

Engler, A. (2022). *The AI Bill of Rights makes uneven progress on algorithmic protections*. Brookings Institution. https://policycommons.net/artifacts/4140999/the-ai-bill-of-rights-makes-uneven-progress-on-algorithmic-protections/4949604/

European Court of Auditors. (2024). *EU Artificial intelligence ambition: Stronger governance and increased, more focused investment essential going forward* (Special Report 08/2024). https://www.eca.europa.eu/en/publications/sr-2024-08

European Data Protection Board. (2023). *EDPB publishes urgent binding decision regarding Meta* [Press release]. https://edpb.europa.eu/news/news/2023/edpb-publishes-urgent-binding-decision-regarding-meta_en

European Digital Rights, Access Now, Algorithm Watch, Amnesty International, Bits of Freedom, European Center for Not-for-Profit Law, European Disability Forum, Panoptykon Foundation, Homo Digitalis, AccessNow, Fair Trials, Irish Council Civil Liberties, Elektronisk Forpost Norge, & PICUM. (2023). *Civil society calls on EU to protect people's rights in the AI Act 'trilogue' negotiations* [Statement]. https://edri.org/our-work/civil-society-statement-eu-protect-peoples-rights-in-the-ai-act-trilogue-negotiations/

European Parliament. (2023). *Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts* (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD)). https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html

European Parliament. (2024). *Artificial Intelligence Act: MEPs adopt landmark law* [Press release]. https://www.europarl.europa.eu/news/en/press-room/20240308IPR19015/artificial-intelligence-act-meps-adopt-landmark-law

Europol. (2023). *ChatGPT - The impact of large language models on law enforcement* (Tech Watch Flash) [Report]. Europol Innovation Lab. https://www.europol.europa.eu/publications-events/publications/chatgpt-impact-of-large-language-models-law-enforcement

Feng, S., Park, C. Y., Liu, Y., & Tsvetkov, Y. (2023). *From pretraining data to language models to downstream tasks: Tracking the trails of political biases leading to unfair NLP models* (arXiv:2305.08283). arXiv. http://arxiv.org/abs/2305.08283

Garg, Y., Seetharam, K., Sharma, M., Rohita, D. K., & Nabi, W. (2023). Role of deep learning in computed tomography. *Cureus*, *15*(5). https://doi.org/10.7759/cureus.39160

Géron, A. (2019). *Hands-on machine learning with Scikit-Learn, Keras, and TensorFlow: Concepts, tools, and techniques to build intelligent systems* (2nd ed.). O'Reilly Media.

Gkritsi, E. (2024, May 24). Staffing questions swirl around Commision's AI Office. *Euractiv*. https://www.euractiv.com/section/artificial-intelligence/news/staffing-questions-swirl-around-commisions-ai-office/

González Fuster, G. (2014). The right to the protection of personal data and EU law. In G. González Fuster (Ed.), *The emergence of personal data protection as a fundamental right of the EU* (Vol. 16, pp. 213–252). Springer. https://doi.org/10.1007/978-3-319-05023-2_7

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., & Bengio, Y. (2014). *Generative Adversarial Networks* (arXiv.1406.2661). arXiv. https://doi.org/10.48550/arXiv.1406.2661

Gstrein, O. J. (2022). European AI regulation: Brussels effect versus human dignity? *Zeitschrift Für Europarechtliche Studien*, *2022*(4), 755–772. https://doi.org/10.5771/1435-439X-2022-4-755

Gstrein, O. J., & Zwitter, A. J. (2021). Extraterritorial application of the GDPR: Promoting European values or power? *Internet Policy Review*, *10*(3). https://doi.org/10.14763/2021.3.1576

Hacker, P., Engel, A., & Mauer, M. (2023). Regulating ChatGPT and other large generative AI models. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, 1112–1123. https://doi.org/10.1145/3593013.3594067

Harris, D. E. (2023). *Voluntary curbs aren't enough: AI risk requires a binding international treaty* [Opinion piece]. Centre for International Governance Innovation. https://www.cigionline.org/articles/voluntary-curbs-arent-enough-ai-risk-requires-a-binding-international-treaty/

Hartmann, T. (2024, February 28). French MPs voice sovereignty, competition concerns after Microsoft-Mistral AI deal. *Euractiv*. https://www.euractiv.com/section/artificial-intelligence/news/french-mps-voice-sovereignty-competition-concerns-after-microsoft-mistral-ai-deal/

Heikkilä, M. (2023, August 7). AI language models are rife with political biases. *MIT Technology Review*. https://www.technologyreview.com/2023/08/07/1077324/ai-language-models-are-rife-with-political-biases/

Helberger, N., & Diakopoulos, N. (2023). ChatGPT and the AI Act. *Internet Policy Review*, *12*(1). https://doi.org/10.14763/2023.1.1682

Hildebrandt, M. (2023). Artificial intelligence law. In J. M. Smits, J. Husa, C. Valcke, & M. Narciso (Eds.), *Elgar encyclopedia for comparative law* (2nd ed., pp. 139–152). Edward Elgar. https://doi.org/10.4337/9781839105609.artificial.intelligence.law

Hu, K. (2023, February 2). ChatGPT sets record for fastest-growing user base – analyst note. *Reuters*. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/

Hu, M. (2020). Cambridge Analytica's black box. *Big Data & Society*, *7*(2). https://doi.org/10.1177/2053951720938091

Jasserand, C. (2023). Experiments with facial recognition technologies in public spaces: In search of an EU governance framework. In A. Zwitter & O. Gstrein (Eds.), *Handbook on the politics and governance of big data and artificial intelligence* (pp. 315–357). Edward Elgar Publishing. https://doi.org/10.4337/9781800887374.00023

Krempl, S. (2024, May 8). AI Act: Datenschützer wollen KI-Verordnung in Deutschland durchsetzen [AI Act: Data protectionists want to enforce AI regulation in Germany]. *heise online*. https://www.heise.de/news/AI-Act-Datenschuetzer-wollen-KI-Verordnung-in-Deutschland-durchsetzen-9713089.html

Kroet, C. (2024, January 24). Commission sets up AI office as sign-off on rulebook nears. *Euronews*. https://www.euronews.com/next/2024/01/24/commission-sets-up-ai-office-as-sign-off-on-rulebook-

nears

Kundu, A. (2024). The AI Act's gender gap: When algorithms get it wrong, who rights the wrongs? *Internet Policy Review*. https://policyreview.info/articles/news/ai-acts-gender-gap-when-algorithms-get-it-wrong/1743

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*, 436–444. https://doi.org/10.1038/nature14539

Li, S., & Newman, A. L. (2022). Over the shoulder enforcement in European regulatory networks: The role of arbitrage mitigation mechanisms in the General Data Protection Regulation. *Journal of European Public Policy*, *29*(10), 1698–1720. https://doi.org/10.1080/13501763.2022.2069845

Madiega, T. (2024). *Artificial intelligence act* (EU Legislation in Progress) [Briefing]. European Parliament. http://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf

Maham, P., & Küspert, S. (2023). *Governing general purpose AI: A comprehensive map of unreliability, misuse and systemic risks* [Policy brief]. Stiftung Neue Verantwortung. https://www.interface-eu.org/publications/governing-general-purpose-ai-comprehensive-map-unreliability-misuse-and-systemic-risks

Mantelero, A. (2022). Human rights impact assessment and AI. In A. Mantelero & B. Data (Eds.), *Beyond data: Human rights, ethical and social impact assessment in AI* (pp. 45–91). T.M.C. Asser Press. https://doi.org/10.1007/978-94-6265-531-7_2

McNeilly, M. (2023). *Will generative AI disproportionately affect the jobs of women?* [Report]. Kenan Institute of Private Enterprise. https://kenaninstitute.unc.edu/kenan-insight/will-generative-ai-disproportionately-affect-the-jobs-of-women/

Milmo, D., & Stacey, K. (2023, November 2). Five takeaways from UK's AI safety summit at Bletchley Park. *The Guardian*. https://www.theguardian.com/technology/2023/nov/02/five-takeaways-uk-ai-safety-summit-bletchley-park-rishi-sunak

Mökander, J., & Floridi, L. (2022). From algorithmic accountability to digital governance. *Nature Machine Intelligence*, *4*, 508–509. https://doi.org/10.1038/s42256-022-00504-5

Morgan, F. E., Boudreaux, B., Lohn, A. J., Ashby, M., Curriden, C., Klima, K., & Grossman, D. (2020). *Military applications of artificial intelligence: Ethical concerns in an uncertain world* [Report]. RAND Corporation. https://www.rand.org/pubs/research_reports/RR3139-1.html

Mukherjee, S., & Vagnoni, G. (2023, April 28). Italy restores ChatGPT after OpenAI responds to regulator. *Reuters*. https://www.reuters.com/technology/chatgpt-is-available-again-users-italy-spokesperson-says-2023-04-28/

Organisation for Economic Co-operation and Development. (n.d.). *Artificial intelligence*. https://www.oecd.org/digital/artificial-intelligence/

Pasquale, F. (2015). *The black box society: The secret algorithms that control money and information*. Harvard University Press. https://doi.org/10.4159/harvard.9780674736061

Perez, F., & Ribeiro, I. (2022). *Ignore previous prompt: Attack techniques for language models* (arXiv:2211.09527). arXiv. https://doi.org/10.48550/arXiv.2211.09527

Petkauskas, V. (2023, April 6). ChatGPT tied to Samsung's alleged data leak. *Cybernews*. https://cybernews.com/news/chatgpt-samsung-data-leak/

Prior, B. (2023, July 9). South African lawyers use ChatGPT to argue case – Get nailed after it makes up fake info. *MyBroadband*. https://mybroadband.co.za/news/software/499465-south-african-lawyers-use-chatgpt-to-argue-case-get-nailed-after-it-makes-up-fake-info.html

Roberts, H., Cowls, J., Hine, E., Morley, J., Wang, V., Taddeo, M., & Floridi, L. (2023). Governing artificial intelligence in China and the European Union: Comparing aims and promoting ethical outcomes. *The Information Society*, *39*(2), 79–97. https://doi.org/10.1080/01972243.2022.2124565

Russel, S., Perset, K., & Grobelnik, M. (2023, November 29). Updates to the OECD's definition of an AI system explained. *The AI Wonk*. https://oecd.ai/en/wonk/ai-system-definition-update

SAS. (n.d.). *SAS documentation*. SAS Support. https://support.sas.com/en/documentation.html

Schertel Mendes, L., & Kira, B. (2023). The road to regulation of artificial intelligence: The Brazilian experience. *Internet Policy Review*. https://policyreview.info/articles/news/road-regulation-artificial-intelligence-brazilian-experience/1737

Shah, D. (2023, May 25). The ELI5 guide to prompt injection: Techniques, prevention methods & tools. *Lakera – AI Security Blog*. https://www.lakera.ai/blog/guide-to-prompt-injection

Smuha, N. A., Ahmed-Rengers, E., Harkens, A., Li, W., MacLaren, J., Piselli, R., & Yeung, K. (2021). *How the EU can achieve legally trustworthy AI: A response to the European Commission's Proposal for an Artificial Intelligence Act* [Submission to public consultation]. LEADS Lab @ University of Birmingham. http://dx.doi.org/10.2139/ssrn.3899991

Sparkes, M. (2023, November 2). What did the UK's AI Safety Summit actually achieve? *New Scientist*. https://www.newscientist.com/article/2400834-what-did-the-uks-ai-safety-summit-actually-achieve/

Stegmaier, P., Visser, V. R., & Kuhlmann, S. (2021). The incandescent light bulb phase-out: Exploring patterns of framing the governance of discontinuing a socio-technical regime. *Energy, Sustainability and Society*, *11*. https://doi.org/10.1186/s13705-021-00287-4

Taylor, L. (2023). Can AI governance be progressive? Group interests, group privacy and abnormal justice. In A. Zwitter & O. Gstrein (Eds.), *Handbook on the politics and governance of big data and artificial intelligence* (pp. 19–40). Edward Elgar Publishing. https://doi.org/10.4337/9781800887374.00011

TensorFlow. (2023). *TensorFlow guide* [Guidance]. https://www.tensorflow.org/guide

Thrun, S. (1998). Lifelong learning algorithms. In S. Thrun & L. Pratt (Eds.), *Learning to learn* (pp. 181–209). Springer. https://doi.org/10.1007/978-1-4615-5529-2_8

van der Beek, P. (2023, January 16). AP wordt landelijk coördinator algoritmetoezicht [AP to become national algorithm surveillance coordinator]. *Computable*. https://www.computable.nl/artikel/nieuws/overheid/7459315/250449/ap-wordt-landelijk-coordinator-algoritmetoezicht.html

Veale, M., & Zuiderveen Borgesius, F. (2021). Demystifying the Draft EU Artificial Intelligence Act – Analysing the good, the bad, and the unclear elements of the proposed approach. *Computer Law Review International*, *22*(4), 97–112. https://doi.org/10.9785/cri-2021-220402

Weiser, B. (2023, May 27). Here's what happens when your lawyer uses ChatGPT. *The New York Times*. https://www.nytimes.com/2023/05/27/nyregion/avianca-airline-lawsuit-chatgpt.html

Whiddington, R. (2023, April 17). A photographer submitted an A.I.-generated image to a prestigious art competition to be 'cheeky.' It won a top prize anyway. *Artnet News*. https://news.artnet.com/new

s/boris-eldagsen-photography-award-sony-ai-generated-images-dall-e-2286622

Wiggers, K., Corrall, C., & Stringer, A. (2023, July 31). ChatGPT: Everything you need to know about the AI-powered chatbot. *TechCrunch*. https://techcrunch.com/2023/08/02/chatgpt-everything-you-need-to-know-about-the-open-ai-powered-chatbot/

Zellinger, P. (2023, December 10). Das umfangreichste KI-Gesetz der Welt und seine Lücken [The world's most comprehensive AI law and its loopholes]. *Der Standard*. https://www.derstandard.at/story/3000000198931/das-umfangreichste-ki-gesetz-der-welt-und-seine-luecken

Zenner, K. (2023, July 20). A law for foundation models: The EU AI Act can improve regulation for fairer competition. *The AI Wonk*. https://oecd.ai/en/wonk/foundation-models-eu-ai-act-fairer-competition

Published by

Alexander von Humboldt Institute for Internet and Society

in cooperation with

CREATe

cnrs centre internet et societe

R&I IN3 Internet interdisciplinary Institute Universitat Oberta de Catalunya

University of Tartu Johan Skytte Institute of Political Studies