SCHRIFTEN DES IRDT

TRIER STUDIES ON DIGITAL LAW

Raue / von Ungern-Sternberg / Kumkar / Rüfner (ed.)

Artificial Intelligence and Fundamental Rights

The AI Act of the European Union and its implications for global technology regulation

Volume 4

Raue / von Ungern-Sternberg / Kumkar / Rüfner (ed.)

Artificial Intelligence and Fundamental Rights

The AI Act of the European Union and its implications for global technology regulation

TRIER STUDIES ON DIGITAL LAW

Published by: Verein für Recht und Digitalisierung e.V.
Institute for Digital Law Trier (IRDT)

Trier, 2025
Volume 4

SCHRIFTEN DES IRDT

TRIER STUDIES ON DIGITAL LAW

Published by Verein für Recht und Digitalisierung e.V.,
Institute for Digital Law Trier (IRDT), Behringstr. 21, 54296 Trier, Deutschland

The German National Library lists this publication in the German National Bibliography; detailed bibliographic information is available at http://dnb.d-nb.de
This book is also available as E-Book at https://www.epubli.com/shop

ISBN: 9783565013197
URN: urn:nbn:de:hbz:385-20241022094426034-3932285-6
DOI: https://doi.org/10.25353/ubtr-dab1-9b5c-1ec6

The Trier Studies on Digital Law are sponsored by Trier University and the IRDT.

UNIVERSITÄT TRIER

IRDT Institute for Digital Law Trier

Content

Introduction

Artificial Intelligence and Fundamental Rights
Antje von Ungern-Sternberg 1

Chapter 1: Making of the AI Act

The AI Act- brief introduction
Irina Orssich 7

From Definition to Regulation: Is the European Union Getting AI Right?
Joanna J. Bryson 11

Chapter 2: Prohibited AI Practices

Prohibited AI Practices under the EU AI Act
Patricia García Majado 35

Chapter 3: High-risk AI Systems

Risk Narrative: Deconstructing the AIA's Risk-Based Approach as a Regulatory Heuristic
Tobias Mahler 57

Data Governance under the AI Act
Lea Ossmann-Magiera, Lisa Markschies 75

Human Oversight under the AI Act and its interplay with Art. 22 GDPR
Tristan Radtke 91

The Regulatory Approach of the European Union's Artificial Intelligence Act
David Restrepo Amariles, Aurore Troussel 111

Chapter 4: Brussels Effect? The outside perspective

The US Perspective
Margaret Hu 129

AI Governance and Asia Aspect
I-Ping Wang 135

List of Abbreviations 157
Index of Authors 159

Introduction

Artificial Intelligence and Fundamental Rights

Antje von Ungern-Sternberg

Artificial Intelligence applications often intersect profoundly with fundamental rights. In response to this, the EU AI Act[1] aims to erect safeguards for fundamental rights while providing a legal framework that stimulates innovation. The annual conference of the Digital Law Institute Trier, held at the Electoral Palace in Trier on September 26 and 27, 2024, explored whether that delicate balancing has been successful.

I. EU Regulation on AI: The Rights-Driven Approach

One might say that the EU's new AI Act is a typical European endeavour. Europeans are not only proud of European culture, universities, or landscapes, they also believe in the power of law. Particularly in the field of digital technology, the EU has been very active in creating new laws intended to protect values like fair competition, fundamental rights, rule of law, and democracy, as evidenced by the General Data Protection Regulation, the Digital Markets Act and the Digital Services Act – and now by the AI Act. It is true, AI promises great potentials and opportunities: It will make services more accessible and efficient and thus make life and work easier, it will boost the economy, and it will increase public health and safety, to name only the most obvious hopes. However, one can also fear the negative impact of AI on fundamental rights, rule of law and democracy, which led the EU legislator to pass the AI Act. This piece of legislation fits nicely into the distinction made by *Anu Bradford*, contrasting three models of regulating digital technologies: the Chinese state-driven, the US-American marked-driven and the European rights-driven approach.[2] But EU regulation in the digital sphere is

[1] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).

[2] Cf. *Anu Bradford*, Digital Empires, 2025, p. 105 ff.

also criticised from within the EU. Some fear that too much red tape, too many cumbersome duties concerning risk management, data governance, technical documentation, record-keeping and transparency will seriously hamper entrepreneurs, particularly small and medium size enterprises and start-ups. Therefore, regulating AI as well as interpreting and applying the new AI Act is a difficult task. It is important to discuss how duties under the AI Act can be understood to accommodate legitimate concerns and to avoid undue burdens upon developers, distributers, and deployers of AI. If Europeans have learned one lesson from data protection law in EU, i.e. from the General Data Protection Regulation, it is the lesson that some of its obligations are (construed) too broadly.

This book focuses on AI and fundamental rights. This is an obvious choice, as AI techniques have tremendous implications for human rights. The EU and its predecessors, the EEC and the EC, have a long history of commitment to fundamental rights, even if a formal rights catalogue – the EU Charter of fundamental Rights – was only proclaimed in 2000, and became binding only in 2009.[3] Interestingly, over the last ten to fifteen years, the European Court of Justice (ECJ) has considerably strengthened and expanded its focus on fundamental rights. One could say that the ECJ has developed from a guardian of the internal market and the market freedoms to a guardian also of fundamental rights.[4]

One of the Court's important tasks is to in ensure that EU legislation is interpreted and implemented in conformity with fundamental rights.[5] Due to the hierarchy of norms, this rule of interpretation and implementation applies to laws which do not mention fundamental rights at all. But it applies all the to more to laws like the AI Act, which repeatedly refers to fundamental rights in its recitals and the enacting terms[6] as this demonstrates the legislator's intention to protect fundamental rights trough the legislation in questions. In the AI Act, this intention is made very clear in Art. 1 which defines the purpose of the Act to be "to improve the functioning of the internal market and promote the uptake of human-centric and trustworthy artificial intelligence (AI), while ensuring a high level of protection of health, safety, *fundamental rights enshrined in the Charter, including democracy*, the rule of law and environmental protection,

[3] Art. 6 (1) [1] EU Treaty.

[4] For an overview, see *Robert Schütze*, An introduction to European law, 3rd ed., 2020, p. 83 seqq.

[5] Cf. Art. 51 (1) Charter of Fundamental Rights of the European Union.

[6] Cf. Recitals 1, 3, 5, 6, 7, 8, 9, 10, 17, 20, 22, 28, 32, 34, 43, 46, 48, 52, 53, 57, 58, 59, 60, 65, 66, 67, 70, 72, 75, 77, 91, 92, 93, 96, 118, 121, 139, 140, 155, 157, 171, 176 AI Act, and Art. 1 (1), Art. 2 (4), Art. 3 (49) (c) and (65), Art. 5 (2), Art. 6 (3), (6), (7) and (8), Art. 7 (1) (b), (2) (e) and (i), and 3 (a) and (b), Art. 9 (2) (a), Art. 10 (2) (f) and (5), Art. 13 (3) (b) (iii), Art. 14 (2), Art. 27, Art. 28 (7), Art. 36 (7) (e), (8) (a) and (9) (a), Art. 40 (3), Art. 41 (1) (a) (iii), Art. 43 (6) AI Act – to list only the explicit references to "fundamental rights" in the first for chapters of the AI Act.

against the harmful effects of AI systems in the Union and supporting innovation" (emphasis added). Thus, the crucial question is: How far are the rights and obligations under the AI Act shaped and influenced by fundamental rights? This volume is intended to contribute to this question.

II. AI and Fundamental Rights: Issues

AI technology and applications raise a wide range of fundamental rights questions, from human autonomy to discrimination or data protection. This brief introduction cannot deal with all of them, but it can touch upon three important issues. The first is functionality. If AI systems are developed based on wrong assumptions or models, or if they are trained with faulty or biased sets of data, they do not work properly. In this case, a multitude of fundamental rights ranging from life, bodily integrity, or property to autonomy and equality are at stake. That's why AI Act classifies high-risk AI system very broadly. Apart from autonomous cars and other areas of harmonised legislation on product safety, it covers AI systems listed in Annex III. These systems have in common that they are used in situations where individuals' rights need to be particularly protected, first, because individuals are faced with powerful actors such as law enforcement agencies, migration control authorities, and employers, or, second, because individuals need a certain infrastructure like education, essential services, elections, or judiciary to be able to exercise their rights. Several provisions in the AI Act are intended to ensure the proper functioning of AI applications. Notably risk-management (Art. 9 AI Act), data governance (Art. 10 AI Act), technical documentation (Art. 11 AI Act), and record-keeping (Art. 12 AI Act) are important duties to establish and keep corresponding standards. Standards for training data and machine learning, for example, help to avoid that discrimination embedded in our current society and current data is perpetuated into the future.

Another important challenge is that AI makes surveillance and manipulation easy and widely accessible. AI allows and facilitates to collect data, to identify people by their biometric data (face, voice, key stroke, gait), to track, classify, rate and rank people, to create profiles, to find weak spots and to exploit them, and to create fake news, fake accounts, deep fakes. The AI Act bans some of these techniques under Art. 5, i.e. subliminal techniques, social scoring, criminal risk assessments, emotion inferences, biometric identification and classification systems. And it imposes transparency obligations in Art. 50, for examples the duty to label social bots or deep fakes. These provisions help to guarantee fundamental rights like data protection and to privacy, but they also protect human autonomy – i.e. freedom from deception and manipulation – as a

precondition for other freedoms like free speech and free elections. Democracies must find ways to separate facts from fakes, to distinguish the truth from lies, otherwise the very foundation of democracy is at stake.

Finally, if human decision-making and human activities can increasingly be replaced by AI, what role should remain for humans? Which tasks can be left to AI, which tasks should be reserved for humans? Art. 14 AI Act demands that high risk AI systems are "effectively overseen" by humans. This important duty will help AI systems to be accepted and trusted by those who are hesitant. And it is an important endeavour to establish where human oversight is grounded in fundamental rights and – to go one step further – to ask which tasks will have to be performed, not only overseen by humans in the future. I assume that fundamental rights prevent that some tasks are replaced by AI, for example human communication with a judge in the courtroom (based on judicial rights) or democratic decisions-making (based on democratic rights).

III. Outline of the Book

The book starts with the technical foundations of AI regulation and their relevance for safeguarding fundamental rights. First, *Irina Orssich* who has been coordinating AI policies at the EU Commission for a considerable time, i.e. notably in DG Connect and at the newly established AI Office, gives a "Brief Introduction to the AI Act". Orssich who has been involved in the AI Acts' legislative process and is now responsible for the Commission's implementation sums up the core of the Act. Her introduction is followed by an assessment of the AI Act by *Joanna Bryson*, Professor of Ethics and Technology at the Hertie School of Governance, intitled "From Definition to Regulation: Is the European Union Getting AI Right?". Bryson gives insights into her experience with the drafting process as an academic and treats the topic of definitions. Her overall conclusion is that the EU's regulatory approach is to be welcomed.

The next section of the book analyses the expansive scope of the EU's AI Act, the regulation pyramid with the differentiated approach distinguishing between prohibited AI practices, high-risk AI systems and other AI applications, and assess its impact on fundamental rights. *Patricia García Majado*, a postdoctoral researcher in law at the University of Oviedo, deals with "Prohibited AI Practices Under the EU AI Act". García Majado looks into the legislative history and the scope of Art. 5 AI Act. In her view, Art. 5 pursues laudable intentions balancing fundamental rights and technological progress, but has also some weaknesses, notably since the prohibitions are vague and have many exceptions. *Thomas Mahler*, a professor of law the University of Oslo, treats high-risk AI systems. His article "Risk Narrative: Deconstructing the AIA's Risk-Based

Approach as a Regulatory Heuristic" deconstructs the AI Acts' so-called 'risk based approach'. Mahler argues that while the AI Act's reliance on risk is genuine, it does not amount to a singular, cohesive methodology. Rather, it comprises a patchwork of multiple legislative strategies, each engaging with the concept of risk in distinct and sometimes inconsistent ways, thereby offering a flexible framework intended to tailor regulatory obligations to the level of risk posed by AI systems.

The following chapters are dedicated to diligence obligations introduced by the AI Act and their pivotal role in preserving fundamental rights, especially data governance, human oversight and the important role of self-assessment. *Lea Ossmann-Magiera* and *Lisa Markschies*, who are both researchers in law at Humboldt-Universität Berlin, cover "Data Governance under the AI Act" and examine the data quality requirements under Art. 10 AI Act. According to the authors, this obligation is a crucial one, but it is not yet clear the extent to which it will be able to solve pressing problems such as biased data and data contamination. *Tristan Radke*, a postdoctoral researcher in law at TU München, treats "Human Oversight under the AI Act and its interplay with Art. 22 GDPR". His chapter analyses the classification of human oversight measures under the AI Act and argues that – even though the concept of human oversight is challenging – the different approaches under the AI Act and Art. 22 GDPR may work well together. Nevertheless, Radtke finds that the effectiveness of human oversight depends in particular on effective tools in practice. Finally, *David Restrepo Amariles*, Associate Professor of Artificial Intelligence and Law at the École des hautes études commerciales Paris (HEC), deals with "The Regulatory Approach of the European Union's Artificial Intelligence Act", which rests – to a great extent – on self-assessment. He claims that the AI Act aims to balance technological innovation and fundamental rights protection by relying heavily on self-assessment and on co-regulation tools, like sandboxes. According to *Restrepo Amariles*, this approach appears to offer flexibility to providers and deployers, even if it increases their compliance burden and may raise concerns about enforcement consistency and the effectiveness of compliance mechanisms.

The last panel of the conference discussed the possible global implications of the AI Act and a possible 'Brussels effect'. Three legal scholars from outside Europe contributed to the panel and gave an American, an Asian, and an African perspective on AI regulation. We are grateful that *Margaret Hu*, Professor of Law and Director of the Digital Democracy Lab at William & Mary Law School, and *Wang I-Ping*, Professor of Law at National Taipei University, Taipei University, found the time to put down their views in writing. Margaret Hu shows that the U.S. has fallen behind in regulating AI and that the rapid advancement is starting to run at odds with fundamental rights. Wang I-Ping demonstrates how China, Japan, Singapore, and Taiwan focus on promoting AI technology development, often guiding industry practices through non-legally binding guidelines or frameworks.

Chapter 1: Making of the AI Act

The AI Act- brief introduction[1]

Irina Orssich

The AI Act[2] entered into force on 1 August 2024. It lays down harmonised rules for development, placing on the market, putting into service and use of Artificial Intelligence (AI) in the Union.[3] Its aim is to promote innovation in - and the uptake of AI, while ensuring a high level of protection of health, safety and fundamental rights in the Union, and of democracy and the rule of law. This means that there is now for the first time a regulatory framework balancing the potential benefits of AI with the need to protect citizens against its potential downsides.

The AI Act follows a risk-based approach, introducing rules for the development and use of AI systems that are commensurate to the risks, and do not go beyond what is necessary.

- A small number of uses of AI that are considered a clear threat to fundamental rights of people and that are listed in Article 5 AI Act, will be banned. This includes for example AI that manipulates human behaviour to circumvent users' free will, social scoring, or a number of applications in the field of biometrics.

- The core of the AI Act is its approach to AI systems that are considered to be 'high-risk' because they pose particular risks to fundamental rights, health and safety, for example medical devices, or AI used for recruitment. These systems are classified according to Article 6 and listed in Annex I and III AI Act. High-risk systems are subject to a set of obligations and requirements, including for -

[1] The opinions and views expressed in are solely the responsibility of the author and do not represent the official views, position, or endorsement of the European Commission.

[2] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act) (Text with EEA relevance), PE/24/2024/REV/1, OJ L, 2024/1689, 12.7.2024.

[3] Article 1 AI Act.

example data governance, transparency and cybersecurity, as well as certification processes. Standards will play a key role in the implementation of the requirements: they will, if adopted by the Commission, grant a presumption of conformity to the requirements of the AI Act.

- Moreover, certain uses of AI, for example chatbots or deepfakes, are subject to transparency obligations, so that users exposed to those systems are aware that they are confronted with an AI and can take informed decisions. It will also have to be ensured that AI-generated content can be detected as such.

- To ensure trustworthiness of large AI models, so-called 'general-purpose AI models', a two tiered approach was adopted: Light-touch transparency obligations for all general-purpose AI models, and additional obligations related to the management of risks for very capable and impactful models. There will be codes of practice as a central tool to fleshing out the rules, drafted in cooperation with industry and experts. This approach is complemented by a strong governance centred around an alert system, which allows to intervene to quickly address risks as they arise. The regulatory approach on general-purpose AI models has been developed in light of the global discussions, for example on the G7 Principles and Code of Conduct.

- All other AI systems, and that is the vast majority, can be developed and used subject to the existing legislation without additional legal obligations under the AI Act. For those systems providers and deployers may choose to adhere to voluntary codes of conduct.

The European AI Office, established in February 2024 within the Commission, has the mission to reflect the EU AI strategy: implementing the AI Act, fostering innovation in AI and engaging internationally on AI to promote the EU's human-centric approach. It oversees the AI Act's enforcement and implementation together with Member States. The AI Act establishes a two-tiered governance system, where national authorities are responsible for overseeing and enforcing the rules for AI systems, while the EU level is responsible for governing general-purpose AI models.

With regard to the AI Act, the AI Office is:

- Contributing to the coherent application of the AI Act across Member States, including the set-up of advisory bodies at EU level, facilitating support and information exchange;

- Developing tools, methodologies and benchmarks for evaluating capabilities and reach of general-purpose AI models, and classifying models with systemic risks;

- Drawing up state-of-the-art codes of practice to detail out rules, in cooperation with leading AI developers, the scientific community and other experts;

- Investigating possible infringements of rules, including evaluations to assess model capabilities, and requesting providers to take corrective action;

- Preparing guidance and guidelines, implementing and delegated acts, and other tools to support effective implementation of the AI Act and monitor compliance with the regulation.

The AI Act will be fully applicable 2 years following its entry into force, with some exceptions: prohibitions, definitions and provisions related to AI literacy will take effect after six months (2 February 2025), the governance rules and the obligations for general-purpose AI models become applicable after 12 months (2 August 2025) and the rules for AI systems - embedded into regulated products, listed in Annex II, - will apply after 36 months (2 August 2027). In this context, the Commission is promoting the AI Pact, seeking the industry's voluntary commitment to anticipate the AI Act and to start implementing its requirements ahead of the legal deadline. Launched in May 2023, the AI Pact is structured around two pillars: Pillar I focuses on stakeholder engagement. The AI Office holds a series of seminars on different aspects of the implementation of the AI Act. Pillar II consists of voluntary pledges by the industry

From Definition to Regulation: Is the European Union Getting AI Right?

Joanna J. Bryson[1]

I. Introduction

Thank you for the tremendous honour of having me at this meeting, and particularly on this, the first panel, "Introduction and technical foundation", a theme that is about at the limits of my competence for matters of the law. It is just fantastic to be here, and I already have at least five questions for the first and previous speaker, Irina Orssich.

1. Has digital governance already proved a bad idea? (No.)

Let me begin though with some answers. Many are asking what is the cost of the EU's propensity for regulation to its own—and the rest of the world's—economy, and technical competence?

In 2021 I published (with Helena Malikova) a paper called "Is There an AI Cold War?"[2] that showed statistics about AI competence globally. Many beforce us had been comparing only the 'top' few companies by market capitalisation, a disinformative measure that makes those who regulate well (and enforce antitrust) look weak. Malikova and I looked at everyone defending at least two patents in 2019 in one WIPO category that captured quite a lot of AI — an admittedly somewhat arbitrary measure. We looked both at the number of patents and the aggregate market capitalisation of every such company, and we found that for both those measures the EEA (the area impacted by EU digital regulations like the GDPR and the AI Act) was comparable to China. Also, the rest of the world (excluding the US, CN, and EEA) were greater than the EU + CN on both measures. And the US dominated at that time the *entire* rest of the world on both measures. Of course, the US has substantial global dominance also of financial

[1] Hertie School, Centre for Digital Governance.

[2] Joanna J. Bryson and Helena Malikova. "Is There an AI Cold War?" In: Global Perspectives 2.1 (June 2021), p. 24803. issn: 2575-7350. doi: 10.1525/gp.2021.24803. eprint: https://online.ucpress.edu/gp/article-pdf/2/1/24803/480097/globalperspectives_2021_2_1_24803.pdf. url: https://doi.org/10.1525/gp.2021.24803.

instruments and markets used in our measures. In more recent data[3] we see the EU performing even more strongly against China, and some indication that the US market capitalisation dominance may be at least partly a bubble.

I wasn't that surprised by these outcomes. Well, I was by the strength of the "rest of the world", but I shouldn't have been for the same reason I wasn't surprised about the EU. It is difficult to believe any strong economy is fully incompetent at AI, given how much AI and the rest of the digital have been taken up in all sectors. Similarly, increasing importance of the digital economy to many EU member states, also makes it unsurprising that there is a great deal of EU AI competence.

So data indicates the EU is doing a great job of innovation.[4] Yes, there are direct costs of compliance that are evident to companies. Though I hear from companies in other sectors with digital components that the costs of GDPR compliance are laughably trivial compared to their sectoral compliance costs. This fact doesn't help digital companies unfamiliar with compliance and terrified they are doing it wrong. Member states should help such companies, and increasingly many governments are providing public advisory systems to get companies over these initial hurdles. But more importantly, the EU regulations are written to benefit the EU digital markets. The benefits are less transparent, particularly to small and inexperienced companies, but are implicit in the growth of the digital sector. The primary goal of the GDPR was to grow the EU digital economy by making a more uniform single digital EU market. But of course, we needed to do that in a safe way, compliant with member state laws.

2. Roadmap

The impacts of regulation are not the talk I was asked give. I was asked to talk about the definition of AI, and how that impacted the AI regulation, or Act. I was trying to understand why anyone would want me to talk for half an hour about definitions, and don't worry, I decided I wasn't going to. I decided instead to start out discussing the broader context of the AI Act, and then turn to the question of the regulation's legitimacy. Where did the AI Act come from? How did an academic like me get involved? Was I really involved? I wasn't one of the heroes in the room for the final 37 hours of the trilogue, unlike Irina Orssich. But we as academics have other kinds of roles to play.

[3] Wiebke Dorfs and Joanna J. Bryson. "Global Artificial Intelligence Competition: Examining Current State and Drivers". In preparation. 2025.

[4] see also Knut Blind, Crispin Niebel, and Christian Rammer. "The impact of the EU General data protection regulation on product innovation". In: Industry and Innovation 31.3 (2024), pp. 311–351. doi: 10.1080/13662716.2023.2271858. eprint: https://doi.org/10.1080/13662716.2023.2271858. url: https://doi.org/10.1080/13662716.2023.2271858; Tom Wehmeier et al. *The State of European Tech 2023*: *Europe must embrace risk now to shape the future.* https://www.investeurope.eu/media/7424/atomico-state-of-european-tech-report-2023.pdf. 2023.

So I'm going to give you a little history, including of my own involvement, but most importantly to emphasise the legitimacy, transparency, and openness of the legislative effort, which I have been hearing some people attack. I will also get to the definition questions, as well as some other present concerns.

3. EU AI regulatory context

But first I will finish setting the context, with some more words on EU digital governance. There's a lot more going on that affects AI than just the AI Act. Please have a look at Figure 1. The original was again due to my collaborator Helena Malikova; I've added some enhancements and updates.

The General Data Protection Regulation (GDPR) underlies all the other AI regulation. As I mentioned earlier, it really hasn't been bad for the EU, but let's talk about what good it does. First, it creates a single digital market, originally harmonising the laws for 28 states. We don't have this through peer review yet, but I have reason to believe the value of that harmonisation far exceeded the costs of compliance, even for Small to Medium sized Enterprises (SMEs). This surprised the student who did the work a lot — they had wanted to look at costs to SMEs; they didn't expect to find benefits. But again, it didn't surprise me that much, because I remember that growing the EU's digital economy was a big part of the goal of the GDPR. Second, the GDPR addresses the member states' obligations under the Universal Declaration of Human Rights (UDHR) to protect the rights of all humans within their borders. These rights include democracy.

I think of personal data privacy as being sort of like airspace. Before you had airplanes, nobody talked about defending airspace. It is the airplane and now drones that make airspace something we need to defend. And it is AI that makes personal data a hazard, or more of a hazard. You know, the Nazis used personal data (with the help of IBM.) ISIS uses data on human relationships to control villages, but entirely with pencils and paper, because they think the US can penetrate any computer.

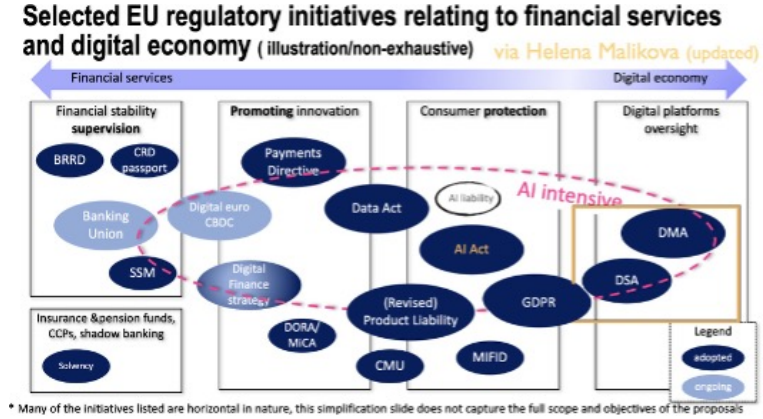


Figure 1: Many EU regulatory initiatives concern AI. This graph is borrowed by permission from its original author, Helena Malikova; the status of the initiatives is updated by Joanna Bryson in late April 2025.

But when you get digital search and synthesis, it gets a lot easier, and more can be done.

Brilliant people noticed this before it happened. There were already discussions in the OECD concerning data protection in 1960s, then an agreed set of principles in 1980 that various governments subsequently used as the basis of laws and binding directives. Just as now, people agreed principles first, and then wrote the early data protection laws. And we had pretty good data protection laws already in the 1980s and on through the 2000s. The GDPR just levelled these up — harmonised them — for the members of the EU, as well as bringing them into the 21*st* Century e.g. with algorithmic transparency requirements.

In some sense, if you can grab our data, you can grab us. You can manipulate us. This is what the Digital Services Act (DSA) is about. It is supposed to make the EU into a place with a safe, predictable, trusted online environment. The DSA deals with profiling, recommendation systems, and targeted advertisements. The elections of 2015 and 2016, Brexit and Trump, made evident the importance of having mechanisms to check whether the GDPR was really being respected. Different people were getting different advertisements online, and these advertisements were not being recorded and disclosed, as was required under at least UK election law. Law was not being enforced in digital space. The DSA mandates that we are provided with enough information to ensure that laws are being respected, and to understand the social outcomes. It is like the GDPR allowing us to do the work we are actually already legally obliged to do.

The Digital Markets Act (DMA) is super important and interesting. Historically, the United States innovated antitrust (market concentration) laws because they were afraid that individuals were acquiring so much power that possibly they could not be

governed by a democracy. This was in the late 1800s, remember the American style of democracy was still an experiment, and off to a rocky start. We were walloped by our colonial power in 1812, and then we had a horrific civil war. Then all this power started concentrating around those dominating new technologies that eased operation of businesses at great distances. So antitrust was innovated in the late 1800s and enforced through about 2000. In fact, the US and UK demanded Germany and Japan also put antitrust in their post World War II constitutions. The allied governments reasoned that if you allow a company to have too much power, either it takes over the government, or the government takes over it, and either way you get autocracy. Writing in the context of 2025, I would now say maybe 'company' should be 'sector.' Just look at various countries with the oil-based 'Resource Curse', or the UK (finance) or the US (tech).

Unfortunately, in 2000 the US largely stopped enforcing antitrust.[5] Earlier, around the time that the Soviet economy plateaued, a new twist was introduced into the theory of antitrust enforcement that claimed size of a company wasn't a problem. Governments only needed to ensure consumer prices were fair. That radical theory broke through into the courts right when Bill Clinton was replaced by George W. Bush, during the determination of penalties for Microsoft.[6]

So there is this big question for those not in the US about what do you do? What do we do if the US no longer enforces its own laws against companies that are affecting us in our countries? The DMA is an innovative, super interesting piece of legislation to address that question. Its novelty lies in how interactive it is. It's a more agile form of legislation, where you can negotiate with the affected companies, and move penalties forwards and backwards as they respond.[7] I was originally sceptical about the need for the DMA, because I thought the EU courts should just be backing up its Directorate General of Competition a little more than they had been doing recently. In that or some other way, a more general problem of EU antitrust enforcement should be solved. In my opinion, a lot of regulatory bad comes from people making the digital sector unduly exceptional. I have since heard from people in the German Justice Ministry that the DMA is seen as sort of a testing ground for a new style of antitrust regulation. So maybe the DMA will ultimately lead to general regulatory improvement in other sectors as well.

Finally, the AI Act. In my opinion, it's super boring. Beautifully boring. We are just saying yes, product law applies to digital products. People used to say, "no, these aren't

[5] Tim Wu. *The curse of bigness*. Columbia Global Reports, 2018.

[6] Carl Shapiro. "Microsoft: A remedial failure". In: *Antitrust Law Journal 75* (2008), p. 739.

[7] If you want to know more about the DMA, Article 19 has been running a series of academic meetings on the Digital Markets Act. https://www.article19.org/issue/digital-markets-act/.

products; they are services. How can software be a product? We build it on top of libraries that keep changing!" This makes no sense. It's like saying, "How could we be responsible for our bridge collapsing? It's not our fault we sourced bad concrete." Of course, in every sector you are responsible for your products, including checking your own supply chain. You have to lock down your libraries, and you have to guarantee that any online learning your product does won't take it out of conformity with the performance guarantees you've made for it.

One thing that's really made me very happy is the revised the Product Liability Directive.[8]It explicitly says that software products are products. You can perform services with AI, and you can perform a service developing AI, but the AI itself is definitely a product. And of course the AI Act talks about what things you shouldn't do with AI in the EU, and how to document the AI for application of product law and defence against liabilities. But basically, although I think it is fantastic that we have the AI Act, I do think it is dull compared to the DMA and the DSA. It just brings responsibility to a tiny corner of the software industry, but that corner is where actions happen without humans intervening — to foreshadow the definition of AI. So that is why we have to be particularly clear.

II. Representing and Intervening

I'll turn now to the representativeness of the EU legislative process, and to academic intervention. I can't talk about that in general, I am just giving you the perspective of one academic, but you'll see that quite a lot of what I've done, other people could easily do as well, and many did.

I came to regulation initially in about 2010, when I got invited by some more politically clued-in people than I was to a UK national 'ethics' event. At the time I identified as someone in a computer science department writing theoretical biology about intelligence, though I do have a PhD in the systems engineering of real-time, 'human-like' AI from MIT in 2001. I was mostly known to the policy people for worrying about transhumanism and people who over-identify with AI, or who encourage others to over-identify with it. But I was also worried about Google.

As an academic, I have wound up in some pretty weird situations. One of these weird situations was the Google Advanced Technology External Advisory Council they cancelled — and note, they have never paid me. I'm also friends with two of the four

[8] What a geeky thing to say.

authors of that parrot paper[9] that got people thrown out of Google. I get that a lot of people hate and fear the company for its size and various other reasons. But the main reason I worry about Google is that I loved Google; well, I loved it as a tool. And most of my friends who worked there seem like good people. But I was still worried about the kind of power it had, and what was going on with that power, and who constrained it. Google's was not a kind of power I'd learnt about in grade school civics. I was also worried about Twitter, which I think we still have a lot of questions to answer about about Twitter, and X. Some people thought Twitter wasn't powerful because it didn't have a large market capitalisation, but it was striking that the people behind the Internet Governance Forum saw its raw political communication power and invited it onto panels and so forth even back in 2018. Twitter was founded in 2006, and by 2012 two-thirds of world leaders of the 193 UN member countries were active on twitter, by 2016 it was 90%.[10] Diplomats and journalists also found using Twitter essential. Though now (due to the way X hacks the algorithms) journalists are no longer as well served for dissemination, diplomats and politicians can continue contacting each other and communicating in a verified way, because they don't use recommenders, they just read everything each other says.

Everyone here is excited that the AI Act is now adopted. Of course. Hurrah! But the questions I had about Google and Twitter were more handled by the DMA and the DSA. I've sometimes honestly thought that the AI Act was sort of like a decoy that was there to distract American companies from the actual regulations that were getting at the harms they were doing.

1. Pre-documents: High Level Expert Group and White Paper

So let's look at some of the interventions I've been able to make. A lot of this comes down to being visible, publishing, talking on social media, giving talks in policy contexts whenever people ask. I often felt unqualified to talk, but I knew women decline too many invitations so I would always try to say 'yes' if I could. I learnt that even if the main topic is something I know nothing about, like nuclear war or competition policy, the reason I was invited was I would turn out to be one of only two or three people who knew anything about AI there, and having AI expertise mattered.

[9] Emily M. Bender et al. "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" In: *In Conference on Fairness, Accountability, and Transparency (FAccT '21)*. Canada: ACM, 2021, p. 14.

[10] Matthias Lüfkens. "Guest post: Twitter has become the channel of choice for digital diplomacy worldwide". In: *twitter blog (2016).* url: https://blog.twitter.com/en_us/a/2016/guest-post-twitter-has-become-the- channel-of-choice-for-digital-diplomacy-worldwide.

The first piece relevant to the AI Act I wrote was a blogpost, "A smart bureaucrat's guide to AI regulation"[11] I'd been invited to speak at the ICT Conference by the EU's Austrian presidency in December 2018. I'm not sure who invited me, but I was on some panel, and afterwards I ran into a guy and there was a woman just haranguing him. So it turned out he was Pekka Ala-Pietilä, the head of the AI High Level Expert Group, and the woman was also on the group. A lot of people were frustrated that the HLEG had too much industry, and where they had academia or civil society, they didn't know that much about really building AI. So she was saying, why don't you get people like that woman on that panel (me) that actually know about AI? Then they saw me, and he asked me to do a two-page bullet list for him. And so I wrote it, but I also put it on line as my blog. And I didn't put his name on it then, but I've since heard that he liked telling people he was the smart bureaucrat of the title. And a lot of the content there did wind up influencing the high level expert group's final document.

Then a call went out (Irina Orssich just talked about this process.) Everyone was asked, as you heard. Anyone could put in their commentary. And I did. And again, I blogged my commentary as well.[12] So I just kind of transparently blog all this stuff in case someone else can use it, and to increase the chances I get credit for my ideas. The same thing happened when we got the commission's AI white paper, and I put in comments on that too.[13] In each of these blogs, I include my annotated version of the paper they distributed, as well as whatever comments or forms they have you fill in.

In one of those attachments, I did say something about definitions. I said something I was taught as a first year undergraduate: don't waste your time arguing about definitions. It turns out this is in the CIA handbook since like the 1950s: if you want to slow or stop progress, you get people to argue out definitions. And apparently the first 11 months of the 12 months the OECD had with their two representatives of each member nation, when they were supposed to come up with their finalised AI principles, people were indeed stopping progress by arguing about definitions of AI. The current OECD definition is great, incidentally. We'll come back to definitions and the OECD principles later.

2. First Draft

At this point Brexit finally 'got done,' five years after the (ill-conceived, advisory) referendum. Brexit day, 1 February 2020, was my first day of work at Hertie School of

[11] https://joanna-bryson.blogspot.com/2019/01/a-smart-bureaucrats-guide-to-ai.html.

[12] My comments/critiques on the EU's High Level Expert Group on AI's "ethical guidelines" https://joannabryson.blogspot.com/2019/02/my-commentscritiques-on-eus-high-level.html.

[13] Regulating AI as a pervasive technology: My response to the EU AI Whitepaper Consultation https://joanna- bryson.blogspot.com/2020/06/regulating-ai-as-pervasive-technology.html.

Governance, in Berlin. So I've worked in the EU continuously since October 2002, even if I somehow lost my citizenship rights. As it turned out, one of the people in Hertie School's second cohort was significantly involved in German AI policy, a man named Michael Schönstein. He read that I was coming to Berlin, contacted me, and took me out to talk over lunch within a week or two of me starting my job. He also wound up taking all these documents I produced with him to the negotiations. I don't know if the commission really read all 1500 of the submissions they got, but several people definitely read mine. So Germany basically backed my recommendations. And that was great. When the first draft of the AI Act came out from the commission, I was pretty happy with it. For example, the stuff about product liability. It's not like these were my own ideas, I was just amplifying and filtering everything I heard at all these events, at talks, and what I read on Twitter. A lot of the filtering was just saying, you know, this is true or that is not true. That is not how we actually build software, the data doesn't actually program itself into a machine. Humans always architect the system, select the data, algorithm, parameters, tuning procedures, tests. Telling people things like that which unfortunately industry was lying about, and lawyers didn't know much about.

So that was all great, but then Germany had this thing called an election. It keeps happening, every few years you get this new government. So sorry to the non-Germans for a little "inside baseball" here. Schönstein worked for the ministry in charge of labour and society (BMAS), which is why his interests and mine were well aligned — we both thought from the perspective of workers and other citizens and residents, the employed and the unemployed. We want productive and safe societies. In 2020, German AI policy was run across three ministries, the other two being economics and science / research. The national government was a coalition government, and the BMAS was run by the centre left (SPD) who actually got more power after the 2020 election. But somehow the small, libertarian (more than liberal, FDP) party got put in charge of digital policy, and someone in charge hated what I was saying. Maybe Google was buying them a lot of dinners. Google bought me and a bunch of other people a very nice dinner in Berlin in 2023 to tell us about generative AI, in a hotel. Anyway, the head of digital policy and I had a 90 minute phone conversation about definitions. In particular, all the obligations of the AIA only concern 'high risk' AI. A Googler on a panel I was once on[14] claimed only a machine learning algorithm that is fewer than 12 months old was high risk. I said no, the highest risk AI we have had so far was of liking kittens, and it

[14] Organised by the think tank of the EPP — the coalition in the European Parliament of center right parties.

was at least four years after we started liking kittens that Cambridge Analytica actually made that dangerous.

Anyway, back to this unnamed (here) German bureaucrats: after listening to me for 90 minutes he said, "you can't fool me, I have an undergraduate education in computer science. I know for a fact machine learning is different from regular computer science." True, but not relevant. I don't care if the decision system was put together using clockwork and hamsters, if someone gets denied a loan they should have enough explanation to ensure their rights weren't violated.

It became evident that I wasn't going to be influencing German policy during the 'Traffic Light' coalition the same way I had during the previous government. Fortunately, about the same time, I had run into this woman at an academic policy event, a Finnish woman, Meeri Haataja, who is amazing.[15] So she and I coauthored an article that was a response to the European Commission's first AI Act draft.

Now as an academic, you usually try to put work into the best journal it can get into. But if you want to intervene and help legislators, you have to get out a little faster. So Haataja and I published in what was basically a newsletter that had approached me asking me to put something there. And I said, okay, it's not really peer-reviewed, but at least we got this nicely formatted document we could share, "Reflections on the EU's AI Act and How We Could Make It Even Better".[16] We did have some concerns about that first AIA draft, but we were more concerned by the people pointing out these evident weaknesses and trying to tear the whole project down. So we spent more than half the article praising what we thought needed to be kept in it. Meeri also worked out the costs of the act, which the newsletter refused to publish because the article was too long, so we just disseminated those over socarxiv, and that's worked well too.[17] *Wired* even covered the costs 'appendix'.

Speaking of *Wired*, they had a new procuring editor, Angela Chen, who liked my work and had been chasing me for content. And so when I had this fight on the phone with the German, and I was really mad, I emailed her and said, OK, I want to write! About the definition of AI. And she's like, okay. And so I just ranted, and she edited and fixed it, and someone else in marketing made up the title, "Europe Is in Danger of

[15] The event was a week long, online, COVID era event with talks and then a two-day writing crunch. Meeri and I 'met' because we both decided to write the Conclusion first. Here too it quickly became evident how complementary (not similar!) we were in terms of writing, politics, and background knowledge, yet aligned in priorities and goals. Writing with diverse coauthors is very helpful in policy.

[16] Meeri Haataja and Joanna J. Bryson. "Reflections on the EU's AI Act and How We Could Make It Even Better". In: *CPI TechREG Chronicle* (Mar. 2022). url: https://www.competitionpolicyinternational.com/reflections- on-the-eus-ai-act-and-how-we-could-make-it-even-better/.

[17] Meeri Haataja and Joanna J. Bryson. *What costs should we expect from the EU's AI Act?* SocArXiv 8nzb4. Center for Open Science, Aug. 2021. doi: 10.31219/osf.io/8nzb4. url: https://ideas.repec.org/p/osf/socarx/ 8nzb4.html.

Using the Wrong Definition of AI".[18] It's a great title; I think it's hysterical. As much as we graduate students in Edinburgh in the 1990s all sat around arguing about definitions of AI in pubs — now to have those crazy conversations become a threat to Europe. But it is.

So many people have thanked me for writing that article, and putting it in *Wired* where everyone could read it. People from the parliament, from the commission, from the Greens, the EPP, all different kinds of parties, Finnish people, Belgian people, Slovenian people, all because it was in *Wired*. So ironically, some small positive thing came out of the FDP being in government. The argument of that article (which I return to below) is: just make the definition general. However we define AI, what really matters for whether we care about product liability is the nature of the impact it might have, not how the system operates. If you do it simply, great, you should easily be able to defend yourself in court if anyone accuses you of negligence or other misconduct. It's such a simple point, I can't remember how we managed to make it 800 words long.

Coming back to that second half of the newsletter article with Haataja, about the costs of compliance. Basically, the commission said there should be no costs. For a well-run company, AI Act compliance should be free, because, basically you should be doing this much diligence anyway. They may have gotten that from me; I was always telling them about revision control for software. This was the original draft of the AIA, before the Parliament tried to shoehorn in quite a lot more extreme compliance costs. Well, anyway, some American think tank that was kind of famous for trying to interfere with European legislation. Michael Veale, I don't know if you know him, but he heard that these were the guys who had made this claim he snorted his drink out his nose. Anyway, that think tank claimed it was going to be more like two million euro per company per year to comply with the AI Act. Well, Meeri Haataja has both run and consulted for software companies, so she actually sat down and figured out how much it would really cost. It was something like between 80,000 and 200,000 euro, depending on the nature of the company. So not zero, but a lot closer to zero than to two million.

3. The Second Drafts

Some months later, we all got the European Parliament's version. And the Council's version and various other versions of various committees. So Meeri and I wrote another article, this time for a respectable academic on-line law journal where again I'd just recently received an invitation to contribute for some reason. This article we called "The

[18] Joanna Bryson. "One Day, AI Will Seem as Human as Anyone. What Then?" In: *Wired* (2022).

European Parliament's AI Regulation: Should We Call It Progress?".[19] Sometimes American journalists ask me if we were questioning whether the entire AI Act is progress? No, we were talking about, are the second drafts progress past the Commission's original first draft? Again, our conclusion was that there'd been tremendous progress, but also some very problematic insertions. So for example, in the first draft, we were really worried about there being insufficient clarity about the providers, the deployers, and the actual end users — you know, normal people. We were worried about AI being illegal unless it was trained with 'unbiased' data, which fully unbiased data would be noise, it would tell you nothing about the world. This was obviously a 'gotcha' big tech could take to court. Almost everything that we had recommended in our first article was done, there was almost nothing we had to reiterate in the second article. It was incredible. The one exception was that we thought to be truly risk-based there should be a continuum of how much diligence you did, not these absolute risk levels. I had heard Karen Yeung speaking to the Council of Europe, and she said that the amount of diligence you did should be proportionate to your own assessment of risk, which makes sense to me. But as you know, even with the final version, we still have the risk levels. There's another article in this special issue that's all about definitions of risk-based, read that one too. The talk at least was terrific. But the fixed risk levels apparently are what companies prefer, or at least software companies presently think they would prefer. They want certainty about their obligations, though it seems like they may also just be people trying to avoid all regulation by being adequately 'low risk.' Fortunately, the Product Liability Directive now makes it explicit liabilities apply to all software anyway.

So the work the parliament did on their draft was great, but included a few scary additions as well. There were some extremely costly impact assessments being asked for, and it sounds like some of that has stayed in to the final version. Specifically human rights assessments in hiring, but also some of the big sustainability asks, and our question was: what did these have to do specifically with AI? There is of course evidently, definitionally, nothing more important than sustainability. Literally, we will not persist without sustainability, but it is not specific to AI. Most data-centre costs presently seem to be about streaming high definition video, and of course there's encryption costs and cryptocurrency mining. Both of those costs now are higher than AI is even projected to be by 2030.[20] Meeri and I were worried that this was kind of an attack on AI regulation, which would again make people try to avoid the entire label of AI, and then be able to

[19] Meeri Haataja and Joanna J. Bryson. "The European Parliament's AI Regulation: Should We Call It Progress?" In: *Amicus Curiae* 4.3 (2023), pp. 707–718. doi: 10.14296/ac.v4i3.5612. url: https://journals.sas.ac.uk/ amicus/article/view/5612.

[20] Jennifer L. "U.S. Data Centers' Power Demand Surges to 46,000 MW: What's Driving the Growth?" In: *CarbonCredits.com* (2025). Updated: March 3, 2025. url: https://carboncredits.com/u-s-data-centers-power-demand-surges-to-46000-mw-whats-driving-the-growth/.

get unsafe software into decision making spaces. The attacks were made by well-meaning people, like my colleague Lynn Kaack, who is now proud of how much her work disrupted the AIA.

So then the parliament apparently tried to address all these costs by saying the whole AI act only should only be about regulating the few biggest companies and forget the rest. This is just terrifyingly wrong. Small companies can do a lot of harm, look at Clearview AI and what can be done with their face recognition software. The entire AIA was geared to making all products safe, and all applications of AI to government processes as well. Again, pumping in these costs then trying to exclude the majority of corporations from oversight seemed like a clear regulatory interference strategy to Haataja and I.

In the end, the "only very large companies" thing only happened in the domain of generative AI, which there I think it was appropriate. In my opinion, all these concerns about AGI and generative AI were something that got shoehorned into the Act in the mayhem of 2023. People saying 'no one ever thought about a system like chatGPT' — what, you never saw *Star Wars* or *Star Trek*? I for one was certainly thinking about both generative models and natural language systems when I was talking about AI regulation. Anyway, the big companies were most likely trying to build moats,[21] by making it very expensive to do generative AI, but now it is only expensive to do generative AI if you are very large. They deserve those extra costs for forcing the whole GPAI thing on the Act, and costing a lot of legislators a lot of sleep, in my opinion.

III. What Could Possibly Go Wrong?

1. Many Things

So now that the AI Act is done, what could possibly go wrong? Well for one, as I mentioned, right from the beginning, in the first draft, there was something in the text about how machine learning could only be trained on unbiased data. What could that possibly mean? We all know the world is biased, right? The vast majority of biases don't matter to us morally, but they all matter in terms of knowing what is going on. Like if I drop something, say a pencil, it is probably going to fall down. That is a bias, as well as a result of physics. We expect that to happen.

[21] Dylan Patel and Afzal Ahmad. "Google "We Have No Moat, And Neither Does OpenAI": Leaked Internal Google Document Claims Open Source AI Will Outcompete Google and OpenAI". in: *Semi-Analysis* (2023). url: https://semianalysis.com/2023/05/04/google-we-have-no-moat-and-neither/.

There is no data about a world that is unbiased. There is also no data about a world that is totally fair. It is true we need to be worried about biases when we collect or purchase data. We do want to use good, representative data, for example on protected categories. We don't only want to know how medicines work on men or white people, but on all people, for example. For an accurate reflection of the world you have to source data in an unbiased way. But you cannot have unbiased data. When I saw that in the first draft, I thought some corporation had told them to include that, just so that the AI Act would be impossible to enforce. So the Act would just get caught up in court battles for years.

That was something I was worried about even under the Biden administration. This wasn't really intrinsic to the AI Act: I was worried about the enormous amount of effort the United States has put into trying to make sure anything about AI is done at the global level, therefore bypassing localised legislation like the AIA. I was worried even in and before 2024 that these efforts were aimed to reduce the sovereignty of the EU, and of everyone else.[22] Now, they say this is in order to create uniformity of legislation, and sure, it is easier to run transnational companies on a single set of rules. But do we really want uniformity? Different regions of the world have different regulatory capacities, enforcement capacities, threats, neighbours, cultures and histories. And I agree with Kant, from a perspective of regulatory capture, you don't want only one world government. So I think we will probably wind up with eight to twelve different regulatory regions. I mean you don't want 193; you don't want every state in Germany with their own bespoke restrictions. The real power of the Brussels Effect is just getting together a big enough market to justify a certain level of compliance burden. So by that reasoning, smaller regions with smaller variations in their asks may also be OK. I believe a number of regions beyond the EU are building up the capacity to make demands on how they protect their residents. We always mention China, but there are a lot more; for example Brazil, Australia, India, and the African Union. Anyway, one minor positive coming out of Trump's blitzkrieg against the US' own government is that a lot of people are seeing they don't want the US' scale of under-regulation. I found that shift very evident and very reassuring at the February 2025 global AI Action Summit, that was run in Paris.

Another possible problem is lack of enforcement. That has been a big problem with GDPR. We seem to be seeing recently a great decision concerning the Irish Apple Tax case, but that took years. So and I'm still worried. I'm also worried about the people getting hired to the EC's oversight boards. It doesn't so far look like a talent search has

[22] David Backovsky and Joanna J Bryson. "Going Nuclear? Precedents and Options for the Transnational Governance of AI". in: *Horizons: Journal of International Relations and Sustainable Development* 24 (2023), pp. 84–95. url: https://www.jstor.org/stable/48761165.

been done across the planet, which you might expect for world-leading legislation, though I also understand wanting to have local expertise.

There's so much misinformation around regulation. For example, claimed tradeoffs with innovation. There's a lack of perceived legitimacy, cooperation; their's claims of increased costs and decreased effectiveness. People are just constantly running the legislation or a government down. Sitting in Berlin, I often hear people saying two totally contradictory things. First: "you should wait until you understand AI to regulate it." We did. We understand what is happening with recommender algorithms, and so we wrote the DSA. We understand that software, including AI, is a product, and so we wrote the AIA. Second, this claim that regulation limits innovation, yet the same developers then saying how much they look so forward marketing under this EU brand of having the most reliable, human-centred AI. So please don't run things down more than they really need to be run down. And recognise that regulation actually *benefits* innovation, for example by keeping markets diverse,[23] or by pumping money into digital economies.

I'm very, very afraid of the co-option of legitimate digital governance services into illegitimate regimes. Let me both illustrate this and take advantage of it, to also talk about biometrics, and whether or not they should have been fully banned by the AI Act.

Some people don't want any face recognition at all. For example, they don't like that it's used for unlocking phones, or for getting through passport controls. Why aren't these uses forbidden? The reason is because they are not being used to surveil the whole of Europe. When you look at your phone, it doesn't match your face against every human possibly in Europe. It just matches it against the pictures you have willingly uploaded. Personally, I don't do that uploading, but a lot of people do. I don't remember getting asked about whether I wanted biometrics in my passport, so that's a little less consensual, but still technologically it's similar. The question isn't where in the world is every citizen of the EU, but rather, what are the odds that the person at the gate is the person shown on the passport.

China has explicitly said they'd like to know where every person is in their country at all times. I've heard that the UK knows where every car is at all times, I don't know whether that's true, but there are a lot of cameras in the UK and apparently a lot of them look at cars' number plates. But we've decided that this isn't the kind of thing the EU does. At my university, Hertie School, the Workers Council were the people tasked with deciding the exact right angle for all security cameras, in order to protect people's

[23] James Bessen. *The New Goliaths: How Corporations Use Software to Dominate Industries, Kill Innovation, and Undermine Regulation*. Yale University Press, 2022.

privacy when visiting each other's office, but to likely notice who was in a space if something was stolen from it. We spend a lot of time on this kind of thing in Germany, and I think that's good practice.

So the exceptions to the ban on blanket surveillance of faces in the EU is, as far as I understand it, similar to the passports or the phones. It's not that we suddenly start tracking everyone due to some crisis. It's rather that a lot of cameras are looking for the faces of very specific terrorists, or of specific, known, kidnapped children. Presumably a court process has happened to nonconsensually upload pictures of these people into devices that are anywhere.

But here is where we come to the illegitimate use of legitimate technology. If there are cameras everywhere that can be set to look for specific faces, and if we know other countries have technology to track hundreds of millions of people, how can we be certain that no one can hack into our system, or take over our government, and then coopt these resources? Maybe we already *are* being tracked everywhere, and there is only a pretence of ignorance. (Actually, we know we already are, but by our mobile phones.)

These sorts of questions just take governing. We have to inspect these systems, we have to put in safeguards, maybe destructive safeguards. We only have to look at what the US is doing right now with DOGE to violate all kinds of privacy and safety considerations that have been put into place, and they've been doing it *very* quickly. Maybe not entirely effectively or accurately, but if the goal is to instil fear, then arbitrary failures help with that anyway. Incidentally, the Chinese social scoring system apparently entirely failed, no one could keep it running. It was intended to help clear up identity and reduce corruption, and we know it was used as an excuse to block travel for some people and various other types of things. But apparently it's entirely collapsed, like many large government attempts at IT.[24]

But regardless, the issue is whether it is OK to have systems that only work correctly when they are well governed. Is it moral to assume that we will always be vigilant with our democracies, and their cyberescurity? It may very well be that people are right to want to limit what we can do with AI, if those capacities can too easily be coöpted. Which doesn't necessarily mean we have to limit the AI; it may mean we have to increase our capacity to 'burn' the digital bridges if coöption happens. If that's technically an option. Otherwise, maybe limiting AI applications does make sense.

[24] Vincent Brussee. "Is China's Social Credit System As We Know It Dead?" In: *The China Story* (2024). url: https://www.thechinastory.org/is-chinas-social-credit-system-as-we-know-it-dead/.

2. The Military Exception, and Borders

I first learnt about military exceptions when I got pulled into the UK's Robot Ethics meeting in 2010. The meeting was organised by two government research funding agencies, the one for the arts and humanities, and the one for the physical sciences. It originally had no deliverables, but we academics needed to have done something to justify the time, and the result was the five British Principles of Robotics. The first three of these were based on Asimov's laws, but reframed to first of all make them computationally tractable, and secondly to be obligations not of a robot, but of a robot's manufacturer or owner / operator. I felt strongly that we needed to put in place something about transparency and anthropomorphism — not deceiving people into thinking the AI was a person, so that was a fourth principle. Then Lillian Edwards said, well if we're going past three, we should also add in an requirement of attribution, so every robot is someone specific's problem. Personally I thought that wasn't practical, but it was the first of these principles that got turned into law, even in the US.

The British principles are remarkably similar to the OECD principles that came eight years later. But the exception is the first principle, the one that derived from Asimov's first law, that a robot could neither kill a human, nor allow a human through tht robot's inaction to come to harm. The original draft of the principle was "Robots should not be designed solely or primarily to kill or harm humans." But the meeting was a very broad range of British sectors, which included military people. And they said, look, we sell these weapons, it's a major part of our economy. What are you talking about? Any law that's never really going to be applied is a bad law. So we added "except in the interests of national security." At first many of us thought this was some kind of bad compromise, but then we learned.

The reason the military exception makes sense in the British principles is very basic. Policing and civil context do not allow killing except to defend your own life, and robots don't have lives. So of course, robots cannot kill or be used to kill in that context. Military and war is a different context. For whatever reason, we do allow preëmptive killing under the laws of war in some contexts, although of course we never condone wars of aggression. So that better explains the first British principle.

Transnational trade organizations like the EU to date and the OECD do not have military elements. Military matters and enforcement more generally are sovereign to the states. The states are still in the Westphalian system, they still have the monopoly of force. They are just coordinating their trade and sometimes harmonising other laws through these other organisations, but the states are the ones with sovereign capacities of enforcement. I think it's evident that part of the reason that this has been kept very clear and clean for both the EU and the OECD is not only that the states defended their

sovereignty. Rather, the US demanded that military concerns belong in NATO. Since coming to Berlin, I've had Americans from their embassy tell me "we prefer working with organisations of which we are members." So as I mentioned, the OECD AI principles are *remarkably* similar to the British robotics ones, but the main change is that there's no mention killing or national security exceptions here. The first OECD principle is about human centring — it makes the implicit point of the British principles explicit.

Understanding these differences, I'm not as worried about the military exceptions as some of my colleagues. But my students and colleagues at Hertie School have taught me to be very worried about a different exception — the exception being made at the borders. Under the Universal Declaration of Human Rights (UDHR), every part of the world where humans reside should be tiled into states, each responsible for defending the human rights of everyone within its borders. But in present law, there are these weird gaps being opened up at the borders. I would prefer borders to be two dimensional planes that cannot contain people, but right now, they're not. There are not only gaps in the rule of law around the edges of countries, but also in holes within them. For example embassies, but those are technically within some other country. But airports are becoming "borders." US universities now are advising academics against taking even domestic flights, because such weird violations of the law are happening so frequently at US airports. But then, we also know these are happening in the streets. Nevertheless, airports are considered particularly dangerous, at least for academics. I want to note that some awful things happened at US borders under the previous Republican president as well, George W. Bush.

There is another problem concerning borders and AI. The digital era blurs borders, because it makes it easier to operate across them. Even knowing whether or not we're at war is harder. Drones regularly strike outside of any declared war. Which as I just mentioned, declaring war is very important. Are you police or are you military? What powers of enforcement do you have, and when can you kill? Don't forget, even the Vietnam War was never declared by the US. It was called a 'police action.' The US hasn't formally declared war since World War II, although at least unlike Russia you so far don't get thrown in jail for calling the Vietnam War a war. Nevertheless, this is all blurring.

I'm not sure that all breakdowns of border regulation are due to digital issues. Empires didn't used to have very well defined borders, only centres. In periods of extremely high elite inequality such as we are experiencing now — and as we experienced in the very early 1900s, individuals can become more powerful than states, which may make borders less relevant. Around 1900 too, that wealth may have consolidated because of new technologies allowing business to be done by fewer companies operating at greater distance. At that time these technologies were things like rail, telegraph, oil (which is

cheaper to ship than coal.) This oddly does come back to being about AI and definitions, because a lot of problems attributed to AI and the digital are in fact problems shared with other technologies, including petrochemical, pharmaceutical, financial. Even global consultancies and pension funds are compromising sovereignty — and creating mutual shared interests. Which in theory should help with both peace and sustainability, but research is not showing these effects.[25]

IV. Definitions

1. Intelligence and Artifact

OK, so let's do this thing. What is the definition of AI? I've heard full professors of AI say we can't build AI because we don't know what intelligence is. But all you have to do is open a dictionary and there is seven definitions. We know what intelligence is. As I said earlier, I learnt as a first-year undergraduate (thank you, University of Chicago) that what you do when you choose a definition is pick the best available one for the communication task in front of you. Definitions are context-specific facilitation of communication. In the context of the AI Act, the main issues are what harms are forbidden in the EU, and what applications are sufficiently important to people's lives that you really, really care about product liability. Given that these are the main concerns, *how* those harms are produced is less important. What matters is sufficient transparency that we can use the law to improve the system overall.[26]

So the definition of intelligence I prefer is one of the oldest, which incidentally is also the one I was taught as an undergraduate in behavioural sciences at Chicago, *and* as a masters student in AI at Edinburgh. Intelligence is an animal's capacity to adjust its behaviour in accordance with changing conditions. This is due to Romanes.[27] It came out a century before I started my undergraduate degree, when scientists were first thinking of intelligence in non-human, evolutionary contexts.

Words just mean how they used, and I'm not arguing from authority here even though Wittgenstein and Quine spent their careers saying things like that. This is what

[25] Josephine Notaras and Emmy Labovitch. *Progress and priorities: reviewing sustainability in key pension systems.* Tech.rep.Principles for Responsible Investment, UNEP,2024. url: https://www.unpri.org/private-retirementsystems-and-sustainability/progress-and-priorities-reviewing-sustainability-in-key-pensionsystems/12285.article.

[26] Urs Gasser and Viktor Mayer-Schönberger. "Guardrails: Guiding human decisions in the age of AI". in: (2024).

[27] George John Romanes. *Animal intelligence*. London: D. Appleton, 1882.

I got into *Science*. AI contains our biases because it's trained on our language. Our language contains biases because it describes our world and our lived experience.[28] This used to be a contested theory; thanks to AI we were able to show the correlations, and since then other scientists have demonstrated the causal chain. You can change people's implicit biases by changing the kinds of things we read.[29]

So a very useful definition of 'intelligence' is that it is the capacity to do the right thing at the right time. Or put another way, it is the computation of action (or components of action plans) from context. This computation, this physical conversion, requires time, space, and energy, which is why omniscience is impossible and AI has sustainability impacts. And I will take 'artificial' to mean something constructed intentionally, which entails responsibility. This is what underlies all my work in AI 'ethics' and regulation — that we attribute responsibility to agencies we are able to influence through social means such as the law. Which means legal and moral agency, at least for manufactured products, must always rest with human adults. AI objects should not be attributed agency. We do things with AI, 'AIs' are never responsible for actions, not even collaboratively.[30] AI as an active noun should only refer to our discipline, which is constituted of people.

2. Government and Rights

Speaking of constituting, since defining 'AI' is so easy, I'm going to close out this chapter by also discussing 'government' and 'rights'.

Once upon a time, we thought of government as someone riding up on a horse to your village, holding a big sword, and saying OK, I have good news and bad news. The neighbours aren't going to be raiding you any more, but now you have to pay me tax. There was some type of implicit or explicit social contract, with some level of coercion by the powerful and consent by the governed.

---

[28] Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. "Semantics derived automatically from language corpora contain human-like biases". In: *Science* 356.6334 (2017), pp. 183–186. issn: 0036-8075. doi: 10.1126/science.aal4230. eprint: http://science.sciencemag.org/content/356/6334/183.full.pdf. url: http://science.sciencemag.org/content/356/6334/183.

[29] Gary Lupyan and Molly Lewis. "From words-as-mappings to words-as-cues: the role of language in semantic knowledge". In: *Language, Cognition and Neuroscience* 34.10 (2019), pp. 1319–1337. doi: 10.1080/23273798.2017. 1404114; Dermot Lynott et al. "Are you what you read? Predicting implicit attitudes to immigration based on linguistic distributional cues from newspaper readership; a pre-registered study". In: *Frontiers in Psychology* 10 (2019), p. 842.

[30] Katie D. Evans, Scott A. Robbins, and Joanna J. Bryson. "Do We Collaborate With What We Design?" In: *Topics in Cognitive Science* n/a.n/a (2023). doi: https ://doi.org/10.1111 /tops.12682. eprint: https ://onlinelibrary.wiley.com/doi/pdf/10.1111/tops.12682. url: https://onlinelibrary.wiley.com/doi/abs/10.1111/tops.12682.

This idea got replaced during the Enlightenment by the idea that power involves some kind of passive obligations to which we're all entitled — discoverable rights. "We hold these truths self-evident, that all men are created equal, endowed by their creator. . . " That's from the US Declaration of Independence. But the thing is, the US system, even version 2.10 of our constitution with its Bill of Rights, doesn't have positive rights. The rights in the American constitution are all negative rights, just the government leaving you alone. So you can assemble, pray, own guns, espouse opinions, whatever. There's an infinite number of things any government can *not* do. But the far more recent UDHR contains positive rights, and these create tensions. It takes money to give people healthcare. And housing and food and even jobs. All these various obligations cost resources to provide, and every country on the planet has signed up to providing these to every human within their borders, however they got there.

I'm no longer theistic or otherwise supernaturalist, but I still have a problem with rights being an impossible target. Remember what I said earlier in the context of military AI: good laws have to be possible to enforce. I was also again here motivated by arguments with Googlers, who really saw government as some kind of opponent. Government was just trying to stop innovation, or take their money or power. Which is nuts. Every government wants to keep their country strong, wealthy, and secure. Every real government. They would all love to have companies like Google thrive. So why couldn't the Googlers, who are quite smart — why couldn't they see that?

As I was trying to work my way towards a better description of government, to address these concerns, I came across an alternative, and I think better definition of rights. Again, these aren't my ideas, I'm just filtering and amplifying. You can read more about these in Bellamy[31] and Tomkins.[32] Ironically, those are chapters in a book by British people sceptical of human rights as understood by the European Union. Hopefully it will also one day also be in my own book, if I ever stop writing articles like this one.

This new framing is that states are constituted by all the people living within them, and governments are the means we the consituents use to coordinate the construction of public goods. Goods like security, health, education, transportation, etc. To me, this is a much closer model of what we are really doing when we consent to be governed —

[31] Richard Bellamy. "Constitutive Citizenship versus Constitutional Rights: Republican Reflections on the EU Charter and the Human Rights Act". In: ed. by Tom Campbell, Keith D Ewing, and Adam Tomkins. Oxford: Oxford University Press, Dec. 2001. Chap. 2, pp. 15–40.

[32] Adam Tomkins. "Introduction: On being sceptical about human rights". In: ed. by Tom Campbell, Keith D Ewing, and Adam Tomkins. Oxford: Oxford University Press, Dec. 2001. Chap. 1, pp. 1–11.

that consent is already in itself a form of participation. And yes, we are all at least partly responsible for the actions of our states.[33]

In framing, this context, rights become something negotiated and optimized as part of that process of creating governance. Rights are no longer discoverable truths, they are concepts we can innovate. It's understandable that they might conflict, in fact they become a good way to express conflicts of interests. And their relative value may shift as contexts change and tensions force us in different directions. For example, during COVID, our right to free movement shifted temporarily in order defend our right to a working healthcare system, which was under external threat.

I have heard people who knew a lot more about the law than I do being worried that the current framing of rights is being weaponised against the law. Which is ironic since rights were meant to move ethics into a firmer, legal foundation. I'm not a lawyer, but to me this alternative framing of rights and governments might be something that would give us better purchase to defend against many of these assaults on our ability to govern and regulate.

V. Conclusion

In summary of my key points: regulation is not zero sum, nor is it antagonistic to corporations. Regulation helps grow economies, including by making them safer. If Russia and America succeed in further fragmenting the EU, they will also be setting their own markets back — remember the story of Google and the 28 states. We know that we are in a context of radical change, not only because of the climate and other sustainability crises, nor only because of rapidly changing digital capacities, but also because of the enormous challenges involved in governing in an era when some individuals have been allowed far too much power relative to the institutions by which we maintain our agreements. As I mentioned before, the last time we let things get this extreme, we were rewarded with World War I.

Regulation is a process, a process of creating regularity sufficient that we can plan around it — build families and other endeavours. It's like breathing; breathing regulates chemical levels in our bodies and blood streams. It's normal that it has oscillations, also disruptions and recovery.

Nobody should be saying "don't regulate us." What does that even mean? Every tech company that says "don't govern us" is receiving so much largesse from their governments. All these digital companies are getting pumped with money because

[33] Hannah Arendt. "Personal responsibility under dictatorship". In: *Responsibility and Judgment*. Ed. by Jerome Kohn. from 1964. New York: Schocken Books, 2003. Chap. 1, pp. 17–48.

every government wants the digital sector to grow. The right questions to be asking are: how do we set regulation up in a way to make sure that we keep perpetuating ourselves into the future? How do we all live well?

In my opinion, the EU's digital legislation, including, but by no means limited to the AI Act, is a really pretty good set of additions to our capacity for regulation. Provided that we defend it and enforce it. And I sincerely believe that once again, the tech companies contesting these regulations will be some of their many beneficiaries.

VI. Acknowledgements

Thank you to Das Institut für Recht und Digitalisierung Trier (IRDT), for the honour of this invitation, their patience and persistence in seeking this chapter, and most of all for the outstanding, important, and timely meeting. The first section title is indeed a reference to,[34] did anyone guess that? Thanks also to Helena Malikova, Meeri Haataja, and Angela Chen for our collaborations, and Owen Cliffe for helping out with a query on LinkedIn.

[34] Ian Hacking. *Representing and Intervening*. Cambridge University Press, 1983.

Chapter 2: Prohibited AI Practices

Prohibited AI Practices under the EU AI Act

Patricia García Majado

I. Summary of Art. 5 of the AIA

The European Artificial Intelligence Act (AIA) regulates AI practices based on the levels of risk that they pose to the Union's fundamental rights and values such that, as recital 26 of the preamble states, 'the approach should tailor the type and content of such rules to the intensity and scope of the risks that AI systems can generate'. AI practices categorized as posing an unacceptable risk are therefore those that cannot be used without risking fundamental rights and values of the European Union (art. 5 AIA). This is the basis for their prohibition. Recital 28 of the preamble is the first point in the text we are concerned with where it mentions such practices, stating that despite AI's many beneficial uses, 'it can also be misused and provide novel and powerful tools for manipulative, exploitative and social control practices. Such practices are particularly harmful and abusive and should be prohibited because they contradict Union values of respect for human dignity, freedom, equality, democracy and the rule of law and fundamental rights enshrined in the Charter, including the right to non-discrimination, to data protection and to privacy and the rights of the child'.

The European legislators are making a presumption *juris et de jure* with this point, understanding that certain practices—as configured in the Act itself—are not, at present, susceptible to a legally beneficial use in current European democratic systems, or at least a use that is not detrimental to individuals' fundamental rights. To the extent that it is the legislators themselves who have made this assessment of risk, not certain subjects *a posteriori* in a given case, the AIA can be said to have established a top-down approach to risk[1]. Hence, art. 5 AIA in some way outlines the legal frontiers of AI within the Union; something that was certainly considered necessary from both a legal scholarship and institutional point of view from the very moment that the proposed

[1] De Gregorio, G., & Dunn, The European risk-based approaches: Connecting constitutional dots in the digital age, Common Market Law Review, 2022, 473.

Act was published by the European Commission[2]. I use the term legal frontier not only because the article clearly excludes certain practices, but because art. 5 AIA in some way sets out the scope of high risk systems: AI systems that do not fall within the prohibition are often categorised as high risk and therefore subject to special controls and guarantees.

The need for art. 5 AIA was, however, accompanied by deep consideration. As the basic objective is the protection of subjects' fundamental rights and the values of the Union, legislators also strove to avoid excessive, unnecessary prohibitions. The very first section of the preamble indicates that the objective of the Act is both to 'improve the functioning of the internal market by laying down a uniform legal framework' and, 'ensuring a high level of protection of health, safety, fundamental rights as enshrined in the Charter of Fundamental Rights of the European Union (the 'Charter'), including democracy, the rule of law and environmental protection, to protect against the harmful effects of AI systems in the Union, and to support innovation'. And 'more prohibition' does not, at least not necessarily, mean 'better regulation', nor therefore does it mean more effective protection of individuals' fundamental rights. Certain fundamental rights can also be exercised via artificial intelligence (such as, for example, the right to artistic creation), while it can safeguard others (such as AI systems in public safety or in healthcare, etc.). In short, it means finding a delicate balance between two extremes. Art. 5 AIA is the expression of a considered, albeit not always simple, consensus.

The provision concerning us comes into force on 2 February 2025 (art. 113 AIA), which is before the date specified for the Act as a whole to come into force (2 August 2026). Hence, recital 179 of the preamble states, 'while the full effect of those prohibitions follows with the establishment of the governance and enforcement of this Regulation, anticipating the application of the prohibitions is important to take account of unacceptable risks and to have an effect on other procedures, such as in civil law'.

The purpose of the following pages is not to make a point-by-point examination of each prohibited AI practice, but rather to offer a series of general reflections on art. 5 AIA that will help to understand its meaning and highlight potential shortcomings. After analysing the legislative origin of art. 5 AIA, its object (introduction in the market, making available, and use) will be examined, as will the prohibitions unrelated to art. 5 AIA that also affect various AI systems, the different types of prohibitions (absolute and relative) that the article establishes, and finally, the limited restrictive scope of the provision.

[2] Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending certain Union legislative acts, 21 April 2021, available at: https://artificialintelligenceact.eu/wpcontent/uploads/2021/08/The-AI-Act.pdf.

II. The legistlative gestation of Art. 5 AIA: What was there initially and what was not

Precisely because art. 5 AIA expresses—as already noted—the difficult balance between protection of fundamental rights and values of the Union and innovation, it was one of the most hotly-debated provisions of the legislation, and therefore among the provisions with the most changes from its initial wording. The legislative process of art. 5 AIA was, in general terms, progressively restrictive, attempting to provide increasing guarantees. Although various actors were involved in the legislative process, it was the European Parliament[3] that made the most, and the most restrictive, amendments to the original text proposed by the Commission, rather than the Council of the European Union[4]. However, it is also worth noting that the process was influenced by opinions and rulings from various institutional actors, albeit ones without legislative power, such as the European Supervisor of Data Protection, the European Data Protection Board, the Economic and Social Committee, etc., along with other non-institutional actors, basically from the third sector (such as Algorithmic Watch, EDRi, Access Now, etc.), who played an important role in safeguarding the Union's fundamental rights[5].

On the one hand, the Act that was finally passed ended up including more prohibitions than were originally considered. The proposed legislation from the Commission generally prohibited manipulative subliminal AI techniques, those that aim to take advantage of specific vulnerable groups, certain social scoring systems, and certain remote real time biometric identification systems in publicly accessible spaces for law enforcement purposes. The initial proposal did not ban biometric categorization systems, facial recognition databases, emotion recognition systems, or police predictive systems for individuals which were subsequently included—albeit with various modifications and limitations—thanks to various amendments introduced by the European Parliament (amendments 224-227), while the Council only proposed amendments to already established prohibitions, without adding any new bans. The additional prohibitions, however, were also advised or suggested by the European Data Protection Supervisor,

[3] 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)), available at: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.html (last accessed on 2 January 2025).

[4] Council of the EU, Presidency Compromise Text (2021/0106(COD)) (29 Nov. 2021), available at: https://data.consilium.europa.eu/doc/document/ST-14278-2021-INIT/en/pdf (las accessed on 2 January 2025).

[5] See, for example, the report: 'An EU Artificial Intelligence Act for Fundamental Rights. A Civil Society Statement', 2021, signed by 123 European and international organizations. Available at: https://edri.org/wp-content/uploads/2021/12/Political-statement-on-AI-Act.pdf.

the European Data Protection Board,[6] and by the Economic and Social Committee,[7] and also highlighted by other non-institutional actors, who criticized the excessive laxity of the original art. 5 AIA[8].

In addition, this progressive restriction was also apparent in the reinforcement or expansion of already established prohibitions. This is because either the *subjective* scope of the prohibition was expanded or—more commonly—because the *objective* scope was expanded. With regard to the former, see, for example, how in the original Commission text, social scoring systems were prohibited when used by public authorities, while in the final Act, that distinction was removed—as proposed by the Parliament and by the European Council—prohibiting such systems from use by both public authorities and strictly private bodies. This is because private bodies introducing such systems in the market, putting them into service, or use can, in certain circumstances, also harm fundamental rights.

The expansion of the objective scope of the prohibitions may be illustrated by the case of manipulative AI techniques. The Commission proposal prohibited an 'AI system that deploys subliminal techniques beyond a person's consciousness', understood as those that use 'audio, image, video stimuli that persons cannot perceive, as those stimuli are beyond human perception' (recital 29). However, because manipulation is not only at the subliminal (imperceptible) level, but also at the liminal, the final Act also prohibited practices that used 'purposefully manipulative or deceptive techniques, with the objective, or the effect of materially distorting the behaviour of a person or a group of persons by appreciably impairing their ability to make an informed decision (...)'—as proposed by Parliament (amendment 215)—as people in such cases 'can still be deceived or are not able to control or resist them' (recital 29). For instance, consider a chatbot that is used to get people to reveal their passwords.

There was a similar expansion for AI systems that use techniques exploiting people's vulnerabilities. The Commission's proposal prohibited any 'AI system that exploits any of the vulnerabilities of a specific group of persons due to their age, physical or mental disability, in order to materially distort the behaviour of a person pertaining to that group (...)', whereas the vulnerabilities covered by the final text were broadened to include those arising from a 'specific social or economic situation' –'such as persons living in extreme poverty, ethnic or religious minorities', as the preamble states in recital 29—

[6] EDPB-EDPS Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), 18 June 2021, available at: https://artificialintelligenceact.eu/wp-content/uploads/2022/05/AIA-EDPBEDPS-Opinion-18-June-21.pdf. See pp.2 and 3 for proposed prohibitions from these institutions.

[7] Opinion, European Economic and Social Committee, AI Regulation (INT/940 – EESC-2021-02482-00-00-AC-TRA (EN) 5/8) available at: https://artificialintelligenceact.eu/wp-content/uploads/2022/05/AIA-EESC-Opinion-22-Sept-21.pdf. See section 4.8 which specifies the AI practices that they think should be prohibited.

[8] See note 6.

which now seems to be aimed at covering any reason for discrimination (art. 21 CDFUE).

Although the legislative process of art. 5 AIA generally contributed to expanding its object of regulation, it is also important to emphasize that some of the more restrictive amendments to its initial wording, proposed in particular by the European Parliament, were not accepted. This was no doubt on the understanding that they would mean excessive prohibition that would hinder the achievement of certain objectives that were also important for the Union. One such case involves police predictive systems. The European Parliament proposed prohibiting any 'AI system for making risk assessments of natural persons or groups thereof in order to assess the risk of a natural person for offending or reoffending or for predicting the occurrence or reoccurrence of an actual or potential criminal or administrative offence based on profiling of a natural person or on assessing personality traits and characteristics, including the person's location, or past criminal behaviour of natural persons or groups of natural persons' (amendment 224). It was only to be expected that this proscription would be made more flexible bearing in mind the interests of various member states that already used these types of predictive policing tools for security purposes[9]. And in fact, the current prohibition applies only to *individual* predictive policing systems based solely on creating a profile or evaluation of personality traits, which has watered down Parliament's initially proposed ban.

There was a similar process for remote real-time biometric identification systems, as once again, the European Parliament proposed—unsuccessfully—prohibiting all use of such systems without exception in all settings (amendment 220), not only those related to law enforcement. The same was true for Parliament's proposal to also prohibit 'AI systems for the analysis of recorded footage of publicly accessible spaces through 'post' remote biometric identification systems, unless they are subject to a pre-judicial authorisation in accordance with Union law and strictly necessary for the targeted search connected to a specific serious criminal offense as defined in Article 83 (1) of TFEU that already took place for the purpose of law enforcement', which also ultimately failed to be adopted (amendment 227).

Although, as noted above, the legislative process dealt with many of the main defects in the original art. 5 AIA, it is important to mention others that remained, despite—in certain cases—being expressly highlighted by various actors or by legal scholars. In some cases the scope of application that European legislators have finally opted for is questionable. In other words, in some cases in art. 5 AIA it is difficult to understand—or at least difficult to find explanations to judge the reasoning, fundamentally in the preamble, for the same scope prohibiting certain AI systems but not others.

[9] This had already been predicted by some authors such as Presno Linera, La propuesta de Ley de Inteligencia Artificial Europea, Revista de las Cortes Generales, 2023, 81, p.108.

Without attempting to be exhaustive, and solely as an example, we might mention emotion recognition systems, which are prohibited in the workplace and in education. Recital 44, after noting that expression of emotions varies culturally, and even in the same person, emphasizes 'limited reliability, the lack of specificity and the limited generalisability'. It then goes on to explain that, 'considering the imbalance of power in the context of work or education, combined with the intrusive nature of these systems, such systems could lead to detrimental or unfavourable treatment of certain natural persons or whole groups thereof'. That being the case, it is difficult to fathom why this limited reliability is only a valid reason for excluding such systems in employment and education, because if they are not scientifically reliable, or are highly inaccurate, they may produce harmful results in other contexts as well. Furthermore, the imbalance of power that the legislators note as justifying the prohibition—which is a perfectly valid argument, like the previous one—occurs not only in education and employment, but also in other, even more asymmetrical settings, such as migration and law enforcement, contexts that the European Parliament specifically did attempt to include in the prohibition (amendment 226)[10].

Another example could be real-time remote biometric identification systems, which are only prohibited when used for law enforcement purposes[11], with exceptions laid out in art. 5.1h) AIA. Recital 32 in the preamble explains that such systems, in addition to seriously impinging on people's rights and liberties, 'to the extent that it may affect the private life of a large part of the population, evoke a feeling of constant surveillance and indirectly dissuade the exercise of the freedom of assembly and other fundamental rights', also have technical inaccuracies that 'can lead to biased results and entail discriminatory effects', especially 'with regard to age, ethnicity, race, sex or disabilities'. The preamble continues, 'the immediacy of the impact and the limited opportunities for further checks or corrections in relation to the use of such systems operating in real-time carry heightened risks for the rights and freedoms of the persons concerned (...)'.

In this regard, and in line with what was noted previously, we might ask ourselves whether this feeling of mass vigilance and dissuading effect on the exercise of fundamental rights, the discriminatory bias in the results, and the difficulty of making instant corrections do not also mean that real-time remote biometric identification systems

[10] This was also noted by Díaz González, Prohibited Artificial Intelligence Practices (Article 5), in Huergo Lora and Díaz González (Eds.), The EU Regulation on Artificial Intelligence: A Commentary, 2025 (forthcoming); Carlon, Las Administraciones Públicas ante la Inteligencia Artificial, 2025, p. 77. In addition, Smuha, N., Ahmed-Rengers, E., Harkens, A., Li, W., MacLaren, J., Piselli, R., Yeung, K., How the Eu can achieve legally trustworthy AI, LEADS Lab University of Birmingham, 2021, p. 27, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3899991 proposed prohibition, based on those arguments, in law enforcement.

[11] According to recital 46 AIA, 'law enforcement means activities carried out by law enforcement authorities or on their behalf for the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, including safeguarding against and preventing threats to public security'.

pose an unacceptable risk to fundamental rights and values of the Union in settings other than mere compliance with regulations, such as border control and migration, public security, and healthcare. And of course, whether the same risks are not created when such systems are used by autonomous private actors, not only by public authorities[12]. Once again, these risks are not exclusive to a law enforcement setting.

Legal scholars proposed that the prohibition be extended to all situations where there is some kind of coercion of the individual[13], or more ambitiously, in any setting, as the European Parliament proposed (amendment 220), along with the European Data Protection Supervisor[14]. That does not mean that such systems necessarily be prohibited in these arenas, but instead, it underlines the inconsistency of the European legislators, foreseeing an unacceptable general risk of these systems but then limiting the prohibition of them to a single area. Perhaps it would have been useful to explain why its use for certain purposes would be legally prohibited, while its use for others—all those not related to law enforcement—would be subject, where appropriate, to a judgement of proportionality.

III. The object of Art. 5

Despite art. 5 AIA being entitled prohibited AI practices, what it really prohibits are certain *actions* in relation to these systems; and generally three actions: 'the placing on the market', 'the putting into service' or 'the use' of AI systems. Placing on the market means 'the first making available of an AI system or a general-purpose AI model on the Union market' (art. 3.9 AIA). Putting into service refers to 'the supply of an AI system for first use directly to the deployer or for own use in the Union for its intended purpose' (art. 3.11 AIA). However, although in general 'placing on the market', 'putting into service', and 'use' of prohibited AI systems are banned, it is important to note that

[12] In this regard, amongst others, Barkane, Questioning the EU proposal for an Artificial Intelligence Act: The need for prohibitions and a stricter approach to biometric surveillance, Information Polity, 2022, 147, 154. It should be noted that, according to recital 45, 'law enforcement authorities' means 'any public authority competent for the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, including the safeguarding against and the prevention of threats to public security; or any other body or entity entrusted by Member State law to exercise public authority and public powers for the purposes of the prevention, investigation, detection or prosecution of criminal offences or the execution of criminal penalties, including the safeguarding against and the prevention of threats to public security'.

[13] Smuha, N., Ahmed-Rengers, E., Harkens, A., Li, W., MacLaren, J., Piselli, R., Yeung, K., How the Eu can achieve legally trustworthy AI, ob. Cit., pp.25-26.

[14] EDPB-EDPS Joint Opinion 5/2021 on the proposal for a Regulation of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), 18 de junio de 2021, c.32.

for real-time remote biometric identification systems, only the use is banned, not the other actions, meaning it would be possible to place them on the market and put them into service; although it is not clear why this case is more permissive than the others in art. 5 AIA.

In relation to these actions, it seems clear that placing on the market is *ad intra* in nature as it is confined to the European market. The issue may lie with putting systems into service, given that the supply for first use directly to the deployer ('a natural or legal person, public authority, agency or other body using an AI system under its authority except where the AI system is used in the course of a personal non-professional activity', art. 3.4 AIA) is not restricted to the European Union—as nothing is specified. In this way it would seem, according to the legal text, that supply of an AI system *ad extra*—outside the Union; in other words, exporting such a system—may be considered to fall within the definition of putting into service and therefore be prohibited generally by art. 5 AIA.

However, this interpretation of the definition is ruled out by the scope of application of the Act laid out in art. 2 AIA. It only applies to those responsible for deploying AI systems *that are established or located in the Union* but not to those in third countries—which is the previous hypothesis. Therefore, the only putting into service that is subject to the Act is *ad intra*, in other words, by those responsible for deployment of AI systems in the Union. And according to art. 2 AIA, those who are subject to the scope of application of the Act include, among others, 'providers placing on the market or putting into service AI systems or placing on the market general-purpose AI models in the Union, irrespective of whether those providers are established or located' [art. 2.1 a) AIA]—to ensure, states recital 21, 'a level playing field and an effective protection of rights and freedoms of individuals across the Union'; 'deployers of AI systems that have their place of establishment or are located within the Union' [art. 2.1 b) AIA]; as well as 'providers and deployers of AI systems that have their place of establishment or are located in a third country, where the output produced by the AI system is used in the Union' [art. 2.1 c) AIA][15].

The Act is therefore applicable to situations that are linked to the Union, whether by the location of the supplier or those responsible for deployment, or by the effects of

[15] According to recital 22, 'this is the case, for example, where an operator established in the Union contracts certain services to an operator established in a third country in relation to an activity to be performed by an AI system that would qualify as high-risk. In those circumstances, the AI system used in a third country by the operator could process data lawfully collected in and transferred from the Union, and provide to the contracting operator in the Union the output of that AI system resulting from that processing, without that AI system being placed on the market, put into service or used in the Union. To prevent the circumvention of this Regulation and to ensure an effective protection of natural persons located in the Union, this Regulation should also apply to providers and deployers of AI systems that are established in a third country, to the extent the output produced by those systems is intended to be used in the Union'.

using AI systems within it[16]. Hence, AI systems suppliers whose business is exclusively with non-Union states are outside of the scope of application, meaning that it is possible for prohibited systems to be sold by European suppliers to third countries[17] (as long as the results of that export are not used in the Union). In fact, as some authors have already noted, the French firm Idemia/Morpho has sold a facial recognition system to the Shanghai Public Security Bureau and likewise, the Dutch firm Noldus has sold a tool for analysing facial expressions (Facereader) to the Chinese Public Security Ministry[18].

Nonetheless, during the legislative process for the text, parliamentarians insisted on the need to prohibit export of AI systems that were prohibited by the Act[19]. The European Parliament attempted to introduce the following amendment regarding recital 20 (old recital 10): 'In order for the Union to be true to its fundamental values, AI systems intended to be used for practices that are considered unacceptable by this Act, should equally be deemed to be unacceptable outside the Union because of their particularly harmful effect to fundamental rights as enshrined in the Charter. Therefore it is appropriate to prohibit the export of such AI systems to third countries by providers residing in the Union' (amendment 29). In line with that, Parliament proposed that the scope of application of the Act (art. 2.1 AIA) should include 'providers placing on the market or putting into service AI systems referred to in Article 5 outside the Union where the provider or distributor of such systems is located within the Union' (amendment 147). Those attempts, however, were rejected after long negotiation[20]. That being the case, the final option of European legislators, while being less of an obstacle to European providers' commercial activities in third countries, would also considerably lessen the 'Brussels effect'[21]. This effect is not, or should not be, projecting *ad extra* a merely formal regulatory model, but rather at its core, material protection of fundamental rights,

[16] Ortega Giménez, El ámbito de aplicación territorial del Reglamento de inteligencia artificial, in Cotino Hueso and Simón (Eds.), Tratado del Reglamento de inteligencia artificial de la Unión Europea, 2024.

[17] López Tarruella Martínez, El futuro reglamento de Inteligencia Artificial y las relaciones con terceros Estados, Revista Electrónica de Estudios Internacionales, 2023, 1, 15.

[18] Veale and Zuiderveen Borgesius, Demystifying the Draft EU Artificial Intelligence Act — Analysing the good, the bad, and the unclear elements of the proposed approach, Computer Law Review International, 2021, 97, 101.

[19] This was also a proposal from Cserne, Ducato, Zivkovic, Brown, Couzigou, Leontidis, Oren, Sutherland, Sweeney, Yuksel Ripley, Commentary to the Commission's proposal for the "AI Act" – Response to selected issues, Centre for Commercial Law, School of Law, University of Aberdeen, 2021, p. 4: https://www.abdn.ac.uk/media/site/law/documents/UoA_CCL_response.pdf

[20] Wachter, Limitations and Loopholes in the EU AI Act and AI Liability Directives: What This Means for the European Union, the United States, and Beyond, Yale Journal of Law & Technology, 2024, 671, 681.

[21] Bradford, The Brussels Effect, Northwestern University Law Review, 2012, 107.

ensuring that Europe does not act as an agent of harm to those rights in the global market—and not just in its own market[22]. It is not for nothing that recital 8 in the AIA underscores that this is about promoting 'the European human-centric approach to AI and being a global leader in the development of secure, trustworthy and ethical AI'. However, the extraterritoriality of the regulation occurs when those affected by AI systems are European citizens, rather than in any other case.

In any event, if prohibited AI systems are put on the market, put into service, or used, the administrative fine prescribed is €35,000, or if the offender is a business, 7% of their worldwide annual turnover in the previous financial year, if that is greater (art. 99.3 AIA). In addition, the European Data Protection Supervisor may impose administrative fines on Union institutions, bodies, and agencies of up to €1,500,000 for failure to comply with the prohibition of AI practices (art. 100.2 AIA).

IV. The prohibitions of Art. 5 are not numerus clausus

Although art. 5 AIA stringently establishes certain prohibited AI practices, not all AI practices are prohibited by the article. The prohibitions therefore go beyond this provision—which is in this regard not a *numerus clausus* system—meaning that a more global or harmonized view of regulation on this matter is needed in order to be able to determine what is, or may be, prohibited.

In the first place, one must bear in mind that art. 5.8 AIA states that the prohibitions laid out by the article do not affect others that may come from AI practices infringing other European Union law. This means that an AI practice may be prohibited despite not being within art. 5 AIA if it contravenes some other law, 'including data protection law, non-discrimination law, consumer protection law, and competition law, should not be affected by this Regulation' (recital 45). Therefore, what is legally prohibited without AI is also prohibited when it is used[23]. This makes it clear that the parameters of legality of AI systems are not solely shaped by the Act, but by the rest of Union law. For example, the well-known Spanish supermarket chain, Mercadona, was recently sanctioned by the AEPD (Spanish Data Protection Agency) for using facial recognition

[22] Noted by Almada and Radu, The Brussels Side-Effect: How the AI Act Can Reduce the Global Reach of EU Policy, German Law Journal, 2024, 646, 657. They explain that in order for the Brussels effect to occur, there must be indivisibility of the object of regulation such that it would not be the case if there were AI systems for the European market and different systems created for other jurisdictions. Although in relation to real-time biometric recognition systems for law enforcement purposes, where the prohibition only refers to use, Díaz González maintains the same, Prohibited Artificial Intelligence Practices (Article 5), ob. cit.

[23] Voigt and Hullen, What AI Practices Are Prohibited?, in P. Voigt and Hullen (Eds.), The EU AI Act. Answers to Frequently Asked Questions, 2024, 1, 38.

in some of its stores to prevent people who had committed offences against its employees or property—and who had been found guilty and bound by restraining orders—from entering the store. Use of such a biometric identification system is prohibited based on art. 9 GDPR, and also contravenes other previsions of that legislation[24]. So although the AIA only prohibits real-time biometric identification systems in publicly accessibly spaces for law enforcement purposes (art. 5.1 h AIA), their use by private subjects may be prohibited—as in this case—based on the GDPR.

Finally, it is important to highlight that the prohibited practices are those currently laid out in art. 5 AIA. This may change over time. Art. 112 AIA tasks the European Commission with an annual evaluation of 'the need for amendment of the list set out in Annex III and of the list of prohibited AI practices laid down in Article 5', meaning a review and revision of what systems are considered prohibited in light of technical progress, presenting their conclusions to the European Parliament and the Council. This provision was included thanks to an amendment from the European Parliament as the initial text from the Commission only considered the possibility of review to modify the list in Appendix III (high-risk systems).

This seems clearly necessary given that the regulations are about an area of knowledge that changes extremely rapidly, meaning that periodic review is essential to avoid it becoming obsolete and leading to harm to health, security, and fundamental rights[25]. These are all aspects that, along with advances in the information society, the Commission should take into account when formulating their proposed revisions (art. 112.10 AIA). Hence, practices not prohibited now may become so in the future, perhaps because they do not exist currently, or maybe because their potential harm is unknown or cannot be shown (this is always easier to do once they are put into practice). The opposite may also occur; currently prohibited practices may become permitted if technical progress allows them to be implemented without contravening people's fundamental rights.

While there is a need to avoid regulatory obsolescence weakening protection of people's fundamental rights, security, and health, perhaps it would have been more satisfactory had there been a process allowing the Commission itself to alter the list of AI practices prohibited by art. 5 AIA in concert with other actors. This would have been possible had the Commission been allowed to adopt delegated acts in relation to art. 5 AIA—in the same way it is allowed to modify Appendix II by adding or modifying high-risk AI systems (art. 7 and 97 AIA)—to avoid having to fall back on the ordinary

[24] Proceeding No: PS/00120/2021. May be found at: https://www.aepd.es/documento/ps-00120-2021.pdf.

[25] Legal scholarship has already highlighted this need. See, for example, Smuha, N., Ahmed-Rengers, E., Harkens, A., Li, W., MacLaren, J., Piselli, R., Yeung, K., How the EU can achieve legally trustworthy AI, ob. Cit, pp.221.

legislative process, the lengthy nature of which might provide cover for diminishing legal protection[26].

Lastly, it is important to mention that permitted AI practices (as such, outside the scope of art. 5 AIA) may not be usable if they do not comply with the specific obligations laid out in the Act. Consider, for example, high-risk systems that do not pass conformity assessments (art. 43 AIA). Although the effects may be the same in practical terms—not being able to use the systems—these are clearly very different situations than those in art. 5 AIA. There are some AI systems that by their very nature or purposes/effects are contrary to fundamental rights and values of the European Union, and others that are compatible in principle but cannot be used if they do not comply with certain regulations imposed by the Act. In this latter case, permission to use is related to the guarantees and controls established in the Act meaning that nonconformity could be addressed if these are observed, something that does not happen in the case of art. 5 AIA. In any case, this point serves to illustrate that art. 5 AIA is not the sole method of preventing AI systems from being placed on the market, put into service, or used.

V. Absolute prohibitions vs. Relative prohibitions

Although art. 5 AIA lays out all of the prohibited AI practices, it is important to emphasize that not all of the prohibitions are the same. In some cases—perhaps the minority—the *prohibitions are absolute* in the sense that they prohibit certain systems *per se*, without the ban being affected by the system having certain effects or results. This is the case, for example, of systems for making risk assessments of natural persons, facial recognition databases, emotion recognition systems, and biometric categorization systems. The systems in these cases are prohibited without considering additional variables related to their use or implementation.

In other situations, however, art. 5 AIA sets out what we might call *relative* or *conditional prohibitions* in the sense that they exclude certain AI systems but only in that they produce certain effects or consequences, or have certain specific objectives[27]. This means that the same system may be prohibited or not based on the consequences of its use. This is what happens, firstly, with AI systems that use subliminal, manipulative, or deceptive techniques, which are prohibited if they do so 'with the objective, or the effect of materially distorting the behaviour of a person or a group of persons' [art. 5.1 a)

[26] This was the proposal from Smuha, N., Ahmed-Rengers, E., Harkens, A., Li, W., MacLaren, J., Piselli, R., Yeung, K., How the Eu can achieve legally trustworthy AI, ob. Cit p.21; seconded by (among others) Díaz González, Prohibited Artificial Intelligence Practices (Article 5), ob. cit.

[27] This issue was already highlighted by, among others, Miguez Macho and Torres Carlos, Sistemas de IA prohibidos y sistemas de IA de alto riesgo, in Barrio Andrés, M. et al. (Eds.), El Reglamento Europeo de Inteligencia Artificial, 2024, p.55.

AIA]. The same applies to AI systems that exploit a person or group's vulnerabilities, which are only prohibited if they have this 'objective or effect' [art. 5.1 b) AIA]. It is not necessary, therefore, for them to be used with the intention of altering subjects' behaviours, something which seems to be covered by the terms 'objective' and 'purpose', which indicate a volitional component, and was requested in the original Commission proposal[28]. It is sufficient that this situation is produced, in other words, that the AI system merely produces this 'effect'[29]. The important issue here is that without that purpose or effect, the prohibitions do not operate.

What can happen, however, is that these conditions may be (and in many cases are) cumulative, meaning that a chain of them is needed to trigger the prohibition. In the case of manipulative AI techniques, there is the additional requirement of affecting people, 'causing them to take a decision that they would not have otherwise taken', and that this 'causes or is reasonably likely to cause that person, another person or group of persons significant harm'. For AI systems that exploit vulnerabilities, only the latter condition is laid out. So there is a cumulative requirement of two or three effects: substantial change in behaviour, taking a decision that they otherwise would not have taken (only in the case of manipulative techniques), and causing or being reasonably likely to cause significant harm. A prohibition would only be put into place if all of these conditions were met.

Secondly, another example of relative prohibition may be found in social scoring systems, because, along with the other elements required by [art. 5.1 c) AIA], such systems are only prohibited if the resultant social scoring system causes a certain result: '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'; and/or 'detrimental or unfavourable treatment of certain natural persons or groups of persons that is unjustified or disproportionate to their social behaviour or its gravity'. This means that social scoring systems that classify people based on their behaviour or personal characteristics are not in and of themselves prohibited, but only when they cause this detrimental or unfavourable treatment. They would, for example, be permitted for classifying a worker using data related to a given

[28] Art. 5.1 a) RIA, in the Commission proposal, state: 'the placing on the market, putting into service or use of an AI system that deploys subliminal techniques beyond a person's consciousness in order to materially distort a person's behaviour in a manner that causes or is likely to cause that person or another person physical or psychological harm'. This is highlighted by, for example, Veale and Zuiderveen Borgesius, Demystifying the Draft EU Artificial Intelligence Act — Analysing the good, the bad, and the unclear elements of the proposed approach, ob. Cit. p.99; Nikolinakos, N.T., EU Policy and Legal Framework for Artificial Intelligence, Robotics and Related Technologies - The AI Act, 2023, pp.376-377.

[29] Fernández Hernández, C., Capítulo II. Prácticas de IA Prohibidas, in Barrio Andrés (Ed.), Comentarios al Reglamento Europeo de Inteligencia Artificial, 2024.

employment relationship (the same context) as long as they do not cause disproportionate unfavourable treatment.

Lastly, real-time remote biometric identification systems [art. 5.1h) AIA] would also be subject—according to this argument—to relative prohibitions. And unlike the previous cases, what is applied according to the conditions is the exception to the prohibition. In other words, such systems are prohibited 'unless and in so far as such use is strictly necessary for one of the following objectives', which are set down in the clauses of art. 5.1 h) AIA. So this means that real-time remote biometric identification systems are prohibited *if they are not in pursuit* of one of the legally established objectives. The prohibition is affected in any case, albeit negatively.

The conditional or relative prohibitions make it easier to see the risk-based focus that runs through the Act. There are some practices that are prohibited because the unacceptable risks to the Union's fundamental rights and values are only seen if certain effects/purposes occur—which increase the risks exponentially—otherwise they are not. This may only be assessed on a case by case basis. To put it another way, this shows a 'graduated' legal response—accompanying the risk—that is not apparent in absolute prohibitions (by their very nature), which only offer a single response to what may be a range of variables in one practice.

Establishing these conditional types of prohibition poses added difficulties. Firstly, it gives a wide margin of discretion to the bodies that apply the Act. They may, among other things, feel obliged to perform some kind of *judgement of reasonability or suitability* to substantiate the link between the specific AI practice and the corresponding consequences the Act lays out (effect, harm, etc.) of its use, putting into service, or introducing into the market. However, the main difficulty would be in determining the concurrency of the conditions when in many cases they are not events or circumstances that have happened (*ex post* conditions). In many cases they may be intentions—that the AI system is used with a certain 'purpose' or 'objective'. In other cases they may be possibilities or risks (*ex ante* elements), for example causing 'or being reasonably likely to cause' harm, etc. The concurrence of such elements—without a factual basis—is much harder to confirm, necessitating preventive or probabilistic judgements, which seem to be prone to greater levels of interpretability.

VI. The limited scope of Art. 5 AIA

Despite art. 5 AIA covering a broad catalogue of prohibited practices, as we have emphasized above, it is in fact less restrictive than it might seem at first glance. In the first place, its limited scope is because the prohibitions have a scope of application which is generally relatively narrow. On the whole, various elements need to occur together cumulatively for the prohibitions to be activated. These are sometimes consequences or

effects produced by the AI systems. However, many other times they are objective elements, such as the system having certain characteristics or operating in certain contexts. This clearly makes practical application of such cases more difficult because it needs the successive concurrence of various factors which in some cases are not easy to prove. If only one of them is absent or unproven, the prohibition will not be applicable.

For example, art. 5 AIA does not make a general prohibition of predictive policing systems for individuals. It prohibits (1) an 'AI system for making risk assessments of natural persons'—which excludes systems assessing crime risk by area or locations; (2) 'in order to assess or predict the risk of a natural person committing a criminal offence'—which excludes the use of these mechanisms for investigating an existing offence (*ex post*), i.e., for investigative purposes, as well as excluding administrative infractions; (3) 'based solely on the profiling of a natural person or on assessing their personality traits and characteristics'—which permits the use of AI systems based on other factors (although the above also apply), such as systems that use 'risk analytics to assess the likelihood of financial fraud by undertakings on the basis of suspicious transactions or risk analytic tools to predict the likelihood of the localisation of narcotics or illicit goods by customs authorities, for example on the basis of known trafficking routes' (recital 42).

Secondly, art. 5 AIA has many exceptions to the different prohibition cases. See, for example, the exception for 'AI systems used to support the human assessment of the involvement of a person in a criminal activity, which is already based on objective and verifiable facts directly linked to a criminal activity', which is set out as an exception for individual predictive policing systems[art. 5.1 d)A]; emotion recognition systems in the workplace and education 'intended to be put in place or into the market for medical or safety reasons' [art. 5.1 f) RIA] 'such as systems intended for therapeutical use' (recital 44); and the exclusion of 'labelling or filtering of lawfully acquired biometric datasets, such as images, based on biometric data or categorizing of biometric data in the area of law enforcement' [art. 5.1 g) RIA].

These exceptions are supplemented by those for real-time remote biometric identification systems in public spaces for law enforcement purposes, which are allowed when necessary for certain objectives laid out in art. 5.1 h) RIA: 'the targeted search for specific victims of abduction, trafficking in human beings or sexual exploitation of human beings, as well as the search for missing persons'; 'the prevention of a specific, substantial and imminent threat to the life or physical safety of natural persons or a genuine and present or genuine and foreseeable threat of a terrorist attack'; and 'the localisation or identification of a person suspected of having committed a criminal offence, for the purpose of conducting a criminal investigation or prosecution or executing a criminal penalty for offences referred to in Annex II and punishable in the Member State concerned by a custodial sentence or a detention order for a maximum period of at least

four years'. In fact, rather than a prohibition, this seems like a detailed regulation of the guarantees such systems must have when they are used for these specific purposes[30], which is in effect what the extended text of art.5 AIA is largely concerned with.

The problem of exceptions is not that they exist, but rather that, as clauses that operate as limitations to subjects' fundamental rights, they must be properly justified: they should be necessary and genuinely meet objectives of general interest recognised by the Union or the need to protect the rights and freedoms of others (art. 52 CFREU). In addition, exceptions should be worded a certain, precise way to satisfy the exigence of being provided by law (art. 52 CFREU). This also avoids ambiguity or overbroad wording which might make an exception into a useful way to circumvent a prohibition. Vagueness of prohibitions may, in short, undermine the rights-based purpose of art. 5 and hence harm the fundamental rights of those involved. For example, in relation to individual police predictive support systems, set out as an exception, it is not clear what level of human intervention is needed for such a system to be permitted. Is one person enough? What must they be doing? This is problematic, additionally bearing in mind automation bias that may reduce human intervention to a mere formality. Something similar may occur for medical, and particularly security reasons, which may support the use of emotion recognition systems in the workplace and in education, as they may be interpreted with different scope, by different actors, in such contexts[31].

Thirdly, it is important to bear in mind that, when defining the scope of application, art. 2 AIA excludes certain cases. Art. 2.3 AIA is particularly important here, stating that it will not apply to AI systems 'where and in so far they are placed on the market, put into service, or used with or without modification exclusively for military, defence or national security purposes, regardless of the type of entity carrying out those activities' (art. 2.3 AIA), in other words public or private. Nor will it apply to AI systems which are not placed on the market or put into service in the Union, where the output is used in the Union exclusively for military, defence or national security purposes, regardless of the type of entity carrying out those activities. The original proposal from the Commission, however, only contained the exclusion for military purposes, not the other two (defence and national security), which were proposed by the Council and ultimately included. This exclusion—referring to national security and defence—was

[30] This was noted by, among others, Smuha, N., Ahmed-Rengers, E., Harkens, A., Li, W., MacLaren, J., Piselli, R., Yeung, K., How the Eu can achieve legally trustworthy AI, ob. Cit, p.26. The limited scope of the prohibition was also noted by Barkane, Questioning the EU proposal for an Artificial Intelligence Act: The need for prohibitions and a stricter approach to biometric surveillance,, ob. Cit., p.153.

[31] Cemalovic, Prohibited Artificial Intelligence Practices according to art. 5 of the European Union's regulation on AI – between the too late and the not enough, International Journal of Law and Information Technology, 2024, 1, 11.

also included in the Framework Convention on Artificial Intelligence (art. 3.2 and 3.4)[32].

Therefore, these prohibited AI practices, if they are used exclusively with these purposes, must end up being permitted. However, where the same systems are used for other purposes as well (law enforcement, public safety, etc.)—so called 'dual-use systems'—then they are subject to the Act. The exception only operates for systems that are introduced onto the market, put into service, or used *exclusively* for the purposes noted above. That exclusivity breaks when initially excluded systems are used temporarily or permanently for other purposes or when such uses occur at the same time, being introduced into the market, put into service, or used for an excluded purpose and for one or more non-excluded purposes.

According to recital 24, this exclusion in relation to military or defence purposes 'is justified both by Article 4(2) TEU and by the specificities of the Member States' and the common Union defence policy covered by Chapter 2 of Title V TEU that are subject to public international law, which is therefore the more appropriate legal framework for the regulation of AI systems in the context of the use of lethal force and other AI systems in the context of military and defence activities. As regards national security purposes, the exclusion is justified both by the fact that national security remains the sole responsibility of Member States in accordance with Article 4(2) TEU and by the specific nature and operational needs of national security activities and specific national rules applicable to those activities'. Nonetheless, art.2.3 AIA specifies that the Act will not affect the competencies of member states in matters of national security, 'regardless of the type of entity entrusted by the Member States with carrying out tasks in relation to those competences'. This means that prohibited AI practices may also be implemented by private actors on behalf of member states who have outsourced national security tasks to them[33]. Although the CJEU has defined what national security is[34], it has been argued that the interpretation of the—already very broad—concept may vary from state to state, and may also be easily confused with public safety (whose activities are subject to the Act), making legal certainty difficult in relation to the scope of application of the exception.

---

[32] For a legal comparison of the two texts, see the work of Presno Linera and Meuwese, La regulación europea de la Inteligencia Artificial, Teoría y Realidad Constitucional, 2024, 131.

[33] Gómez de Ágreda, La exclusión de los sistemas inteligencia artificial de seguridad nacional, defensa y militares del Reglamento y el Derecho aplicable, in Cotino Hueso and Simón (Eds.), Tratado del Reglamento de inteligencia artificial de la Unión Europea, Aranzadi-La Ley, 2024.

[34] It relates 'to the primary interest in protecting the essential functions of the State and the fundamental interests of society and encompasses the prevention and punishment of activities capable of seriously destabilising the fundamental constitutional, political, economic or social structures of a country and, in particular, of directly threatening society, the population or the State itself, such as terrorist activities' (CJEU, C-511/18, La Quadrature du Net and Others v Premier ministre and Others, ECLI:EU:C:2020:791, para 135)

In this regard, art. 2.3 AIA seems to constitute the first—and very large—exception to art. 5 AIA. It is not, therefore, unreasonable to think that the invocation of national security—with the legal problems of interpretation noted above—might serve as a pretext for resorting to using prohibited practices based on a need to safeguard it. Consider, for example, the use of some of these prohibited systems for border control, or others such as general predictive policing systems, which may be more susceptible to being used under such a cover.

However, it is worth bearing in mind that use of potential prohibited AI systems under the protection of art. 2.3 AIA will not take place in a legal vacuum. The CJEU has indicated that 'although it is for the Member States to define their essential security interests and to adopt appropriate measures to ensure their internal and external security, the mere fact that a national measure has been taken for the purpose of protecting national security cannot render EU law inapplicable and exempt the Member States from their obligation to comply with that law'. Therefore, in so far as the pursuit of these purposes involves using AI systems that need, for example, data processing activity that involve entities subject to Union law—such as data being collected by private actors—, then such AI systems, despite falling under the art. 2.3 AIA exception, will be subject to, among other things, European data protection legislation and the EU Charter of Fundamental Rights[35].

Finally, it is also important to emphasise that the less than restrictive scope of art. 5 AIA is because, in certain cases, AI systems that it would prohibit are already prohibited by other provisions in Union law. These provisions are in primary legislation such as, but not exclusive to, the EU Charter of Fundamental Rights, the General Data Protection Act, (EU) Directive 2016/680, and in the European Convention on Human Rights. For example, without being exhaustive, biometric categorization systems used to 'infer their race, political opinions, trade union membership, religious or philosophical beliefs, sex life or sexual orientation' may be considered to already by covered by the prohibition in art. 9.1 GDPR, which prohibits treatment of personal data 'revealing racial or ethnic origin, political opinions, religious or philosophical beliefs, or trade union membership, and the processing of genetic data, biometric data for the purpose of uniquely identifying a natural person, data concerning health or data concerning a natural person's sex life or sexual orientation'. Alternatively, a social rating system that discriminates against a certain group (women, immigrants, etc.) may be prohibited outside of the AIA based on, among other things, art. 14 ECHR and art. 21 CFREU. That is not to say, obviously, that art. 5 AIA is not needed, but rather that perhaps it has created *ex novo* fewer prohibitions in the Union than an initial reading might have suggested.

[35] Korff, Opinion on the implications of the exclusion from new binding European instruments on the use of AI in military, national security and transnational law enforcement contexts, European Center for Not-for-Profit Law, 2022, https://ecnl.org/sites/default/files/202210/ECNL%20Opinion%20AI%20national%20security_0.pdf.

VII. Brief conclusions

Art. 5 AIA aims to resolve the difficult balancing act between the protection of the European Union's fundamental rights and values, and the promotion of technological progress within the Union and around the world. The finally approved text, in contrast to the Commission's original proposal, is more restrictive as it is often broader in the proposed prohibitions' objective and subjective scope, also allowing annual review to prevent legislative obsolescence from weakening the protection of the fundamental rights at stake. It is also the result of a particularly participative process, involving not only European institutional actors, but other non-legislative bodies who, in various ways, attempted to put forward their own proposals, many of which—as we have seen—were ultimately accepted.

Nonetheless, despite the Act's laudable intentions, it does have some weaknesses that may contribute to weakening its attempts at protecting fundamental rights. Despite an undeniably restrictive appearance, the reality is, in practice, less so. Firstly, some of the prohibitions the Act establishes are already covered by different Union legislation, meaning that in some cases, it does not add any additional limitations. What is not allowed without AI is not allowed with AI. Secondly, because export of prohibited AI practices to countries outside the EU is permitted, this considerably reduces the Brussels effect of protecting fundamental rights on a global level. Thirdly, many of the prohibitions either have many exceptions to their application or have cumulative requirements—some of which are very difficult to monitor—that need to exist concurrently for the prohibitions to apply. This concurrency will often be difficult to prove. And this is without forgetting that fact that in general, art.5 AIA is occasionally worded very broadly or in very abstract terms, making it hard to determine what it really covers. This issue will not only need the interpretive efforts of the Commission—who are called on to publish directives on the practical application of art. 5 AIA (art. 96.1b AIA)—but also the bodies that apply the law. Applying the regulation to specific cases will help to more precisely outline its scope of application.

VIII. References

Almada and Radu, The Brussels Side-Effect: How the AI Act Can Reduce the Global Reach of EU Policy, German Law Journal, 2024, 646.

Barkane, Questioning the EU proposal for an Artificial Intelligence Act: The need for prohibitions and a stricter approach to biometric surveillance*, Information Polity*, 2022, 147.

Bradford, The Brussels Effect, Northwestern University Law Review, 2012, 107.

Carlon, Las Administraciones Públicas ante la Inteligencia Artificial, 2025.

Cemalovic, Prohibited Artificial Intelligence Practices according to art.5 of the European Union's regulation on AI – between the too late and the not enough, International Journal of Law and Information Technology, 2024, 1.

Cserne, Ducato, Zivkovic, Brown, Couzigou, Leontidis, Oren, Sutherland, Sweeney, Yuksel Ripley, Commentary to the Commission's proposal for the "AI Act" – Response to selected issues, Centre for Commercial Law, School of Law, University of Aberdeen, 2021, https://www.abdn.ac.uk/media/site/law/documents/UoA_CCL_response.pdf

De Gregorio and Dunn, The European risk-based approaches: Connecting constitutional dots in the digital age, Common Market Law Review, 2022, 473

Díaz González, Prohibited Artificial Intelligence Practices (Article 5), in Huergo Lora and Díaz González (Eds.), The EU Regulation on Arificial Intelligence: A Commentary, 2025 (forthcoming).

Fernández Hernández, C., Capítulo II. Prácticas de IA Prohibidas, in Barrio Andrés (Ed.), Comentarios al Reglamento Europeo de Inteligencia Artificial, 2024

López Tarruella Martínez, El futuro reglamento de Inteligencia Artificial y las relaciones con terceros Estados, Revista Electrónica de Estudios Internacionales, 2023, 1, 15.

Gómez de Ágreda, La exclusión de los sistemas inteligencia artificial de seguridad nacional, defensa y militares del Reglamento y el Derecho aplicable, in Cotino Hueso and Simón (Eds.), Tratado del Reglamento de inteligencia artificial de la Unión Europea, Aranzadi-La Ley, 2024.

Korff, Opinion on the implications of the exclusion from new binding European instruments on the use of AI in military, national security and transnational law enforcement contexts, European Center for Not-for-Profit Law, 2022. https://ecnl.org/sites/default/files/202210/ECNL%20Opinion%20AI%20national%20security_0.pdf

Miguez Macho and Torres Carlos, Sistemas de IA prohibidos y sistemas de IA de alto riesgo, in Barrio Andrés, M. et al. (Eds.), El Reglamento Europeo de Inteligencia Artificial, 2024, 48.

Nikolinakos, EU Policy and Legal Framework for Artificial Intelligence, Robotics and Related Technologies - The AI Act, 2023.

Ortega Giménez, El ámbito de aplicación territorial del Reglamento de inteligencia artificial, in Cotino Hueso and Simón (Eds.), Tratado del Reglamento de inteligencia artificial de la Unión Europea, 2024.

Presno Linera, La propuesta de Ley de Inteligencia Artificial Europea, Revista de las Cortes Generales, 2023, 81.

Presno Linera and Meuwese, La regulación europea de la Inteligencia Artificial, Teoría y Realidad Constitucional, 2024, 131.

Smuha, N., Ahmed-Rengers, E., Harkens, A., Li, W., MacLaren, J., Piselli, R., Yeung, K., *How the Eu can achieve legally trustworthy AI*, LEADS Lab University of Birmingham, 2021, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3899991

Wachter, Limitations and Loopholes in the EU AI Act and AI Liability Directives: What This Means for the European Union, the United States, and Beyond, Yale Journal of Law & Technology, 2024, 671.

Veale and Zuiderveen Borgesius, Demystifying the Draft EU Artificial Intelligence Act — Analysing the good, the bad, and the unclear elements of the proposed approach, Computer Law Review International, 2021, 97

Voigt and Hullen, What AI Practices Are Prohibited?, in P. Voigt and Hullen (Eds.), The EU AI Act. Answers to Frequently Asked Questions, 2024.

Chapter 3: High-risk AI Systems

Risk Narrative: Deconstructing the AIA's Risk-Based Approach as a Regulatory Heuristic

Tobias Mahler[1]

While the European Union's Artificial Intelligence Act (AIA) is presented as a landmark "risk-based" framework, this article argues that its reliance on risk levels functions more as a narrative device than a consistently applied methodology. Although the AIA genuinely engages with the concept of risk, it does so through a patchwork of distinct legislative strategies that often construe and operationalize risk in divergent ways. Examining the AIA's structure reveals that the widely used risk pyramid, while serving a crucial heuristic function in policy communication, only partly reflects the law's underlying legal architecture. The AIA primarily establishes two distinct risk-based categories: high-risk AI systems and general-purpose AI models presenting systemic risk, which operate under different regulatory logics. To understand the AIA's regulatory architecture, the article traces the origins of the risk-based approach and distinguishes between risk regulation and risk-based regulation. It finds that divergent interpretations of risk by lawmakers, enforcement bodies, and regulated entities, coupled with the introduction of general-purpose AI models as a separate regulatory object, undermine the coherence of a singular risk-based framework, leading to what the article terms a "risk-regulatory cacophony."

I. Introduction

The European Union's Artificial Intelligence Act (AIA) is frequently portrayed as a clearly defined "risk-based" framework that classifies AI systems according to varying

[1] Professor, Norwegian Research Center for Computers and Law, University of Oslo. Funding for this work was provided by the Research Council of Norway under the aegis of the VIROS project ('Vulnerability in the Robot Society'; project number 288285). This paper was first presented at the conference Artificial Intelligence and Fundamental Rights (Trier, 2024), organized by the Institute for Digital Law Trier (IRDT). I am grateful to the participants for their valuable feedback and suggestions.

levels of risk.[2] This framing is articulated in Recital 26 AIA, which underscores the importance of a risk-based methodology to ensure proportionate regulation. At its core, the AIA's approach appears to align with a managerial or engineering rationality, whereby distinct risk levels—such as "high" or "medium"—are identified with the aim of mitigating risks and calibrating regulatory obligations in proportion to their severity.

The emergence of this risk-based approach has received much attention in the literature. De Gregorio and Dunn described the AIA's risk-based approach as a top-down regulatory model, contrasting it with the GDPR's bottom-up framework.[3] Their emphasis was on the AIA's original structure, relying on four levels of risk referring to certain AI systems, introduced in a top-down manner by the lawmaker.[4] Gellert also noted that the AIA departs from the GDPR, where the risk-based approach is anchored in the accountability principle under Article 24.[5] This suggests that multiple "risk-based" approaches may coexist in European law, each serving a distinct purpose and offering its own advantages and disadvantages.

While most commentators are positive about the risk-based approach in the AIA, some have questioned its normative foundation and the degree to which the approach indicates that the lawmaker carried out a thorough and systematic assessment of AI risk. For example, Edwards has pointed out that the AIA's categorization of AI systems based on risk is not justified by externally reviewable criteria and should rather be regarded as political compromises.[6] Ebers contends that the AIA is not "truly risk-based",

---

[2] This approach is especially emphasized in communications by lawmakers, e.g. 'EU AI Act: First Regulation on Artificial Intelligence' (Topics | European Parliament, 6 August 2023) <https://www.europarl.europa.eu/topics/en/article/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence> accessed 15 April 2025. However, some of the literature has also posed a strong emphasis on the risk-based approach, see e.g. Giovanni De Gregorio and Pietro Dunn, 'The European Risk-Based Approaches: Connecting Constitutional Dots in the Digital Age' (2022) 59 Common Market Law Review 473; Martin Ebers, 'Truly Risk-Based Regulation of Artificial Intelligence How to Implement the EU's AI Act' [2024] European Journal of Risk Regulation 1; Marco Almada and Nicolas Petit, 'The EU AI Act: Between the Rock of Product Safety and the Hard Place of Fundamental Rights' (2025) 62 Common Market Law Review 85; Tobias Mahler, 'Between Risk Management and Proportionality: The Risk-Based Approach in the EU's Artificial Intelligence Act Proposal' [2022] Law in the Era of Artificial Intelligence 247.

[3] De Gregorio and Dunn (n 2).

[4] ibid 489.

[5] Raphaël Gellert, 'The Role of the Risk-Based Approach in the General Data Protection Regulation and in the European Commission's Proposed Artificial Intelligence Act: Business as Usual?' (2021) 3 Journal of Ethics and Legal Technologies 15, 20; Raphaël Gellert, The Risk-Based Approach to Data Protection (Oxford University Press 2020); Katerina Demetzou, 'GDPR and the Concept of Risk: The Role of Risk, the Scope of Risk and the Technology Involved' in Eleni Kosta and others (eds), Privacy and Identity Management. Fairness, Accountability, and Transparency in the Age of Big Data, vol 547 (Springer International Publishing 2019) <https://link.springer.com/10.1007/978-3-030-16744-8_10> accessed 3 March 2025.

[6] Lilian Edwards, *Regulating AI in Europe: Four Problems and Four Solutions* (Ada Lovelace Institute, March 2022) 11 https://www.adalovelaceinstitute.org/wp-content/uploads/2022/03/Expert-opinion-Lilian-Edwards-Regulating-AI-in-Europe.pdf, p. 11.

including because it fails to weigh the benefits of AI against its harms,[7] while Wachter argues for closing loopholes by considering further harms and risks in the act.[8]
The discourse focuses not only on the classification of AI systems, particularly what systems are high-risk AI, but often adopts a broader perspective on risk and regulation. For instance, when Kaminski argues that the dominant mode of AI regulation is risk regulation, she places less emphasis on the AIA's classification of certain AI systems as "high-risk" and instead focuses on the ex-ante obligations imposed on AI providers and deployers to assess impacts and risks. Rather than treating risk as a static category, this perspective highlights the procedural mechanisms through which AI risks are identified, managed, and mitigated, particularly before deployment.[9] This approach constructs AI risk in a dynamic rather than a purely categorical manner, recognizing that risk assessments are context-dependent and evolving rather than fixed.

In the AIA, this type of risk assessment is embedded in the AI provider's conformity assessment[10] and it is further reinforced by an impact assessment requirement for certain AI deployers.[11] Risk and impact assessments also play a central role in the US AI policy discourse and soft law.[12] Kaminski contends that this emphasis on self-assessment comes at the expense of other regulatory tools commonly found in risk-based governance, such as individual rights and recourse mechanisms.[13] Moreover, although it

[7] Ebers (n 2).

[8] Sandra Wachter, 'Limitations and Loopholes in the EU AI Act and AI Liability Directives: What This Means for the European Union, the United States, and Beyond' (2024) 26 Yale Journal of Law & Technology 671, 716.

[9] Margot E Kaminski, 'Regulating the Risks of AI' (2023) 103 Boston University Law Review 1347.

[10] See Articles 9, 16, 17 and 43 AIA and further Claudio Novelli and others, 'AI Risk Assessment: A Scenario-Based, Proportional Methodology for the AI Act' (2024) 3 Digital Society 13; Jonas Schuett, 'Risk Management in the Artificial Intelligence Act' (2024) 15 European Journal of Risk Regulation 367; Henry Fraser and José-Miguel Bello Y Villarino, 'Acceptable Risks in Europe's Proposed AI Act: Reasonableness and Other Principles for Deciding How Much Risk Management Is Enough' (2024) 15 European Journal of Risk Regulation 431. For an overview cf. EY, Trilateral Research, "A survey of artificial intelligence risk assessment methodologies: The global state of play and leading practices identified" (2022) https://www.trilateralresearch.com/wp-content/ uploads/2022/01/A-survey-of-AI-Risk-Assessment-Methodologies-full-report.pdf.

[11] Article 26 AIA. See further Alessandro Mantelero, 'The Fundamental Rights Impact Assessment (FRIA) in the AI Act: Roots, Legal Obligations and Key Elements for a Model Template' (2024) 54 Computer Law & Security Review 106020.

[12] Kaminski (n 9) 1347; National Institute of Standards and Technology, 'Artificial Intelligence Risk Management Framework (AI RMF 1.0)' (National Institute of Standards and Technology (US) 2023) NIST AI 100-1 <http://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf> accessed 8 October 2024.

[13] Kaminski (n 9) 1387. In particular, she highlights that a strong emphasis on ex-ante risk assessment does not provide for individual rights or mechanisms for individual recourse, thereby limiting the avenues through which affected individuals can challenge or mitigate AI-related risks. However, more recently, the Council of Europe's Framework Convention on Artificial Intelligence combines risk management with

has become clear that the EU will not adopt the proposed AI Liability Directive,[14] the classifications of AI systems under the AIA could have consequences for other legal issues, such as tort liability, that also hinge on risk.[15]

Accordingly, important questions arise concerning the nature and coherence of this risk-based approach: Should the AIA be understood as implementing a single, unified framework for risk-based regulation? Or is the "risk-based" label better seen as a convenient narrative device? Is the notion of AI risk[16] a fixed and objectively determinable category, or does it vary depending on context, stakeholder perspectives, and shifting regulatory goals? Addressing these questions requires a careful clarification of what constitutes a risk-based approach[17] and how it impacts lawmaking, compliance strategies, and supervisory practices under the AIA.

This article argues that while the AIA's reliance on risk is genuine, it does not amount to a singular, coherent methodology. Rather, it comprises a patchwork of multiple legislative strategies, each engaging with the concept of risk in distinct and sometimes inconsistent ways. Although the risk-based approach fulfills important heuristic and communicative functions—projecting an image of rational, modern governance—it remains only partially embedded in the law itself. Beneath the structured narrative lies a fragmented legal reality, in which risks are construed differently by lawmakers, enforcement authorities, regulated entities, and affected individuals, following divergent risk acceptance criteria and management logics.

The structure of the article is as follows. Section II examines the concept of risk as a mechanism for bridging the distinct regulatory logics of product safety and fundamental rights protection. Section III introduces the theoretical distinction between risk regulation and risk-based regulation, providing a foundation for assessing the AIA's approach. Section IV traces the policy motivations and political compromises underlying the adoption of a risk-based framework. Section V analyses the heuristic function of the risk pyramid as a narrative tool rather than a faithful representation of the law. Section VI critiques the extent to which the risk-based structure is reflected in the AIA's legal

---

remedies and procedural safeguards for individual rights, cf. Article 16, Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law, Vilnius, 5. December 2024, Council of Europe Treaty Series - No. 225.

[14] Proposal for a Directive of the European Parliament and of the Council on adapting non-contractual civil liability rules to artificial intelligence (AI Liability Directive), COM(2022) 496 final.

[15] Johanna Chamberlain, 'The Risk-Based Approach of the European Union's Proposed Artificial Intelligence Regulation: Some Comments from a Tort Law Perspective' (2023) 14 European Journal of Risk Regulation 1, 8.

[16] Regarding the notion of AI risk see also Peter Slattery and others, 'The AI Risk Repository: A Comprehensive Meta-Review, Database, and Taxonomy of Risks From Artificial Intelligence' (arXiv, 14 August 2024) <http://arxiv.org/abs/2408.12622> accessed 26 August 2024.

[17] On the risk-based approach more generally see De Gregorio and Dunn (n 2); Gellert, 'The Role of the Risk-Based Approach in the General Data Protection Regulation and in the European Commission's Proposed Artificial Intelligence Act' (n 5); Ebers (n 2); Mahler, 'Between Risk Management and Proportionality: The Risk-Based Approach in the EU's Artificial Intelligence Act Proposal' (n 2).

architecture. Section VII discusses how the regulation of general-purpose AI models disrupts the coherence of the risk-based framing. Finally, Section VIII concludes by reflecting on the benefits and drawbacks of the risk-based narrative and its implications for regulatory clarity and coherence.

II. Risk as a Conceptual Bridge

The concept of "risk" is invoked no fewer than 776 times across the AIA and its annexes, underscoring both its centrality and the multiplicity of its uses within the Regulation. Risk serves as a classificatory device in the tiered risk-based approach, forms the basis for substantive obligations to identify, assess, and mitigate potential harms, and operates as a regulatory threshold for enforcement interventions. Beyond these functional roles, the notion of risk assumes a conceptual significance: it acts as a bridging mechanism that connects two otherwise disparate domains—namely, the technical, preventive rationality of product safety regulation and the normative, rights-based rationality underpinning fundamental rights protection.

The AIA is deeply embedded in product safety regulation while simultaneously aiming to ensure a high level of protection and fundamental rights, as enshrined in the EU Charter of Fundamental Rights.[18] The link between these two perspectives is challenging, as they pertain to very different domains of law, each characterized by respective underlying assumptions, key concepts and rationales. The concept of risk therefore has an important function in the AIA, as it bridges the distance between the two perspectives.

Product safety law has historically focused on preventing physical harm caused by defective or hazardous products, such as machinery,[19] toys,[20] lifts[21] or those related to

[18] Almada and Petit (n 2).

[19] Regulation (EU) 2023/1230 of the European Parliament and of the Council of 14 June 2023 on machinery and repealing Directive 2006/42/EC of the European Parliament and of the Council and Council Directive 73/361/EEC (OJ L 165, 29.6.2023, p. 1), cf. Tobias Mahler, 'Smart Robotics in the EU Legal Framework: The Role of the Machinery Regulation' (2024) 11 Oslo Law Review 1.

[20] Directive 2009/48/EC of the European Parliament and of the Council of 18 June 2009 on the safety of toys (OJ L 170, 30.6.2009, p. 1).

[21] Directive 2014/33/EU of the European Parliament and of the Council of 26 February 2014 on the harmonisation of the laws of the Member States relating to lifts and safety components for lifts (OJ L 96, 29.3.2014, p. 251).

"explosive atmospheres."[22]. Accordingly, it has developed regulations aimed at mitigating risks such as explosions, mechanical failures, and chemical hazards. Its approach is inherently preventive and technical, emphasizing compliance with safety standards before a product enters the market.[23] The underlying assumption is that risk can be identified and risk level (such as high) can be estimated, often based on existing data about previous incidents.

This quantitative, engineering-based understanding of risk is also reflected in Article 3(2) AIA, which defines risk as "the combination of the probability of an occurrence and the severity of the harm." This definition aligns with traditional risk assessment methodologies, where risks are calculated based on failure rates and the typical consequences of accidents. Such an engineering-based reasoning would be highly relevant for assessing, e.g., the problems likely to be caused by an AI-driven surgery robot, which could injure a patient.

By contrast, fundamental rights law operates within a markedly different legal and philosophical framework, centering on the protection of individual autonomy, dignity, and freedoms rather than merely preventing physical harm. This perspective naturally affects how risk is assessed. For instance, in the context of the General Data Protection Regulation, risk has been characterized as "related to *potential negative impact* on the data subject's rights, freedoms, and interests" [24] (emphasis added)—a softer, more qualitative notion that extends beyond easily quantifiable harms.

Fundamental rights frameworks establish **protections** against state or private interference, ensuring that individuals are not subject to discrimination, undue surveillance, or unjustified limitations on their freedoms. Rather than primarily guarding against physical harm, they delineate protected spheres aimed at upholding intangible moral values associated with a person's inherent dignity, rather than primarily shielding individuals from physical harm.[25] The risk that fundamental rights may be violated is therefore assessed not in terms of **technical failure**, but in terms of **structural, procedural, and ethical considerations**—such as the fairness, transparency, and accountability of

[22] Directive 2014/34/EU of the European Parliament and of the Council of 26 February 2014 on the harmonisation of the laws of the Member States relating to equipment and protective systems intended for use in potentially explosive atmospheres (OJ L 96, 29.3.2014, p. 309).

[23] For example, more than a thousand technical standards address the safety of machinery, cf. Mahler, 'Smart Robotics in the EU Legal Framework' (n 19) 3.

[24] Article 29 Data Protection Working Party, "Statement on the role of a risk-based approach in data protection legal frameworks" WP 2018 (2014), p. 3; cf. Claudia Quelle, 'Enhancing Compliance under the General Data Protection Regulation: The Risky Upshot of the Accountability- and Risk-Based Approach' (2018) 9 European Journal of Risk Regulation 502, 505. Although Recital 75 GDPR mentions that "(t)he risk to the rights and freedoms of natural persons, of varying likelihood and severity, may result from personal data processing which could lead to physical, material or non-material damage," the Regulation fails to clearly define the concept of risk, cf. Karen Yeung and Lee A Bygrave, 'Demystifying the Modernized European Data Protection Regime: Cross-Disciplinary Insights from Legal and Regulatory Governance Scholarship' (2022) 16 Regulation & Governance 137, 145.

[25] Yeung and Bygrave (n 24) 143.

AI systems. We do not typically discuss the probability of harm or its severity in fundamental rights contexts as quantifiable metrics. Instead, discussions center on arguments and balances, where moral and legal arguments often interplay.

Via the concept of "risk to fundamental rights," the AIA also links product safety to **ethical principles**, which are closely aligned with fundamental rights **thinking**.[26] When designing a new law on AI, the EU could have chosen to transform some of these ethical principles into **legally binding fundamental rights-based principles**, as illustrated by the Council of Europe's **Framework Convention on Artificial Intelligence, Human Rights, Democracy, and the Rule of Law**.[27] However, the EU instead chose to **anchor the AIA within the existing framework of product safety law, which is grounded in regulated entities' risk management efforts**, thus requiring a conceptual bridge between these fields.

Risk, traditionally viewed in product safety law as a **technical matter**—focused on preventing accidents and ensuring mechanical reliability—shifts in the AIA toward a **human-centred perspective** that incorporates **fundamental rights and societal impact**. This evolution marks a significant expansion of how risk is understood in EU product safety law, moving beyond **harm prevention** to encompass **societal and ethical concerns**, including **bias, discrimination and human autonomy**. Thus, the concept of risk plays a significant role in bridging the conceptual and regulatory gap between the distinct legal domains of product safety and fundamental rights. Its utility extends well beyond the confines of the risk-based approach, which primarily serves to structure differentiated regulatory responses to varying levels of AI-related risks. Before turning to a discussion of the risk-based approach itself, it is necessary first to draw a distinction that will illuminate the broader context: namely, the distinction between risk regulation and risk-based regulation.

III. Risk-based Regulation Versus Risk Regulation

A clearer understanding of the risk-based approach emerges when it is viewed through the lens of regulatory theory, which distinguishes between risk regulation and risk-

[26] Cf. Ethics Guidelines for Trustworthy AI developed by the EU High-Level Expert Group on AI (HLEG). These guidelines emphasized respect for human autonomy, fairness, transparency, and accountability, reflecting a rights-based approach to AI governance.

[27] Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy and the Rule of Law Vilnius, 5. December 2024, Council of Europe Treaty Series - No. 225.

based regulation.[28] Risk regulation refers to interventions—whether legislative or otherwise—aimed at controlling or managing risk.[29] In this sense, there is little doubt that a core objective of the AIA is to regulate and reduce risks associated with AI technologies.[30]

By contrast, the term risk-based regulation carries a different meaning, one that has also evolved over time. A contemporary understanding of *risk-based regulation* emphasizes "the intentional design of regulatory regimes in which the nature, content, and stringency of regulatory requirements vary *in proportion to* the perceived riskiness of the regulated activity" (emphasis added).[31] This conception focuses less on the mere existence of risks and more on calibrating regulatory responses to the severity and likelihood of those risks. However, the meaning of this term has shifted over time. In early literature the term risk-based regulation was ascribed a more limited meaning. It was used to describe the practices of regulatory agencies that prioritize enforcement efforts based on the level of risk posed by specific cases, focusing resources on the most problematic or high-risk issues.[32] For example, the AIA also allows for a degree of risk-based prioritization by market surveillance authorities responsible for its enforcement.[33] However, to minimize terminological ambiguity, this targeted prioritization should be more precisely described as a *risk-based enforcement* approach.[34] The contemporary understanding of the term risk-based regulation is broader and encompasses a variety of regulatory regimes, beyond enforcement. Thus, when the AIA adapts regulatory burdens between high-risk and non-high-risk AI systems, or between AI models with or without systemic

[28] "Regulation" can be defined as a "sustained and focused attempts to change the behaviour of others in order to address a collective problem or attain an identified end or ends, usually but not always through a combination of rules or norms and some means for their implementation and enforcement, which can be legal or non-legal." Julia Black and Andrew Douglas Murray, 'Regulating AI and Machine Learning: Setting the Regulatory Agenda' (2019) 10 European Journal of Law and Technology section 4 <https://www.ejlt.org/index.php/ejlt/article/view/722> accessed 8 October 2024.

[29] Quelle (n 24) 509.

[30] For example, Article 1 AIA clarifies this aim as "ensuring a high level of protection of health, safety, fundamental rights enshrined in the Charter, including democracy, the rule of law and environmental protection, against the harmful effects of AI systems in the Union and supporting innovation" (emphasis added).

[31] Karen Yeung and Sofia Ranchordás, 'Regulatory Compliance, Public Enforcement and Private Certification', An Introduction to Law and Regulation: Text and Materials (Cambridge University Press 2024) 277 <https://www.cambridge.org/core/books/an-introduction-to-law-and-regulation/regulatory-compliance-public-enforcement-and-private-certification/DE198970AD87B1DB049055A6D76AEA36> accessed 26 February 2025.

[32] Robert Baldwin, Martin Cave and Martin Lodge, 'Risk-Based Regulation' in Robert Baldwin, Martin Cave and Martin Lodge (eds), Understanding Regulation: Theory, Strategy, and Practice (Oxford University Press 2011) 281–283 <https://doi.org/10.1093/acprof:osobl/9780199576081.003.0013> accessed 26 February 2025; Julia Black and Robert Baldwin, 'Really Responsive Risk-Based Regulation' (2010) 32 Law & Policy 181.

[33] Cf., e.g., Article 79 AIA.

[34] Quelle (n 24) 510.

risk, as further discussed below, this can be seen as a form or risk-based regulation. In this sense, the risk-based approach of the AIA is an example of risk-based regulation.

The difference between risk regulation and risk-based regulation lies in their respective goals: the former focuses on the *goal of managing risk*, while the latter prioritizes achieving a balanced and *proportionate* regulatory framework. Both goals are relevant but require different mechanisms.

As mentioned, the AIA's attempt to classify of AI systems based on risk level—its purportedly "clearly defined risk-based approach"[35]—would indeed qualify as *risk-based regulation*, as it establishes differentiated regulatory obligations that correspond to assessed risk levels. The irony lies in the fact that, although one would expect a "risk-based approach" to focus on managing or reducing risk, this particular classification system is not primarily geared toward doing so. Instead, it operates as a mechanism to ensure legislative proportionality, matching regulatory requirements to the perceived severity of risks without directly mitigating them.[36] Thus, while the AIA as a whole could be considered an example of risk regulation, this is not necessarily the case for its tiered AI system classification. The latter—despite being labelled a "risk-based approach"—is better understood as a tool for structuring regulatory obligations proportionally rather than as a direct instrument for controlling risk.

Regulatory regimes may, but need not, be explicitly framed as risk-based. While the AIA expressly highlights the risk-based approach in its text, it is important to recognize that other legal frameworks—although not formally presented as such—could similarly be interpreted through a risk-based lens. While graduated speed limits arguably reflect risk-based reasoning, we do not ordinarily describe the shift from a highway to a regular road as a transition from a zone of high risk to one of medium risk. In a similar vein, it would be possible to reconstruct vehicle regulation as a risk-based approach, although it is typically framed in terms of qualitative distinctions rather than in explicit gradations of risk. For example, we could add a risk-based framing to the regulation of vehicles on the road. Certain vehicles—say, heavy military tanks—might be classified as posing "unacceptable risk" and thereby be prohibited for civilian use. More common motor vehicles, such as cars, could be labelled "high-risk," subject to strict registration and safety checks. Low-power vehicles like electric scooters might be deemed "limited risk," requiring only basic controls and subject to specific requirements, for example regarding parking on pedestrian areas. Routine bicycles could then fall under a "minimal risk" label and face minimal oversight. Framing these tiers in risk-based terms can be rhetorically useful, but not necessary. Many legal instruments operate with special categories that could, in principle, have been labelled as "high-risk." Examples include Article 9 of

[35] Recital 26 AIA.
[36] Mahler (n 2).

the GDPR, which identifies certain types of personal data as particularly sensitive, and Annex I of the Machinery Regulation, which lists machinery deemed especially prone to risk. In other words, framing a regulatory regime as a "risk-based" regulatory regime is optional, and the additional structure may be useful for heuristic purposes, but the approach itself follows largely from the need for regulatory proportionality.

IV. The Origins of the Risk-Based Approach

This section tracks the origins of the risk-based approach in the AIA. From the outset of the AIA's drafting, concerns were raised that applying a new regulatory framework to all AI systems—particularly in light of a broad definition of "artificial intelligence"—could impose disproportionate burdens on AI developers and innovators, especially those operating within a highly competitive global landscape dominated by actors from the United States and China.[37] Politically, excessive regulation was and is seen as a risk to innovation, potentially weakening the European Union's position in global technology markets.[38] Economically, the imposition of compliance obligations on low-risk or routine AI applications threatened to generate regulatory inefficiencies.[39]

To address these concerns, EU legislators emphasized that the AIA would introduce a tiered classification of AI systems based on risk levels, thereby exempting or reducing regulatory obligations for applications considered less harmful. This structure—now central to the AIA's risk-based approach—evolved from the simpler dichotomy introduced in the European Commission's 2020 White Paper on Artificial Intelligence, which divided AI systems into "high risk" and "other" categories.[40]

A more differentiated model emerged in subsequent policy discussions, possibly influenced by the German Data Ethics Commission's proposal for a risk-based approach

[37] Some of the negative sentiments about the AI Act have made it into the utterances of AI-based chatbots, who are critical of the Act, cf. Umberto Nizza, 'Assessing the Impact of the European AI Act on Innovation Dynamics: Insights from Artificial Intelligences' 8.

[38] Mario Draghi, 'The Future of European Competitiveness: A Competitiveness Strategy for Europe' (2024) <https://commission.europa.eu/topics/eu-competitiveness/draghi-report_en> accessed 25 April 2025.

[39] The industry has consistently argued for limited regulatory burdens, see Joint Industry Statement on the Implementation of the EU AI Act, 'To Seize AI Potential in Europe, Follow the Draghi Report: Reduce Regulatory Burden, Focus on Enabling AI Uptake and Innovation' (17 January 2025) <https://orgalim.eu/wp-content/uploads/Joint-Industry-Statement-on-the-Implementation-of-the-EU-AI-Act.pdf>.

[40] White Paper on Artificial Intelligence – A European Approach to Excellence and Trust 16, European Commission (February 19, 2020), available at http://ec.europa.eu/info/sites/default/files/commission-white-paper-artificial-intelligence-feb2020_en.pdf (last accessed 12 March 2025), p. 17. See further Mahler (n 2) 250.

to AI.[41] That proposal already featured a multi-level risk pyramid: with no new regulatory measures at the bottom, graduated obligations depending on risk level, and prohibitions for systems deemed unacceptable at the top.

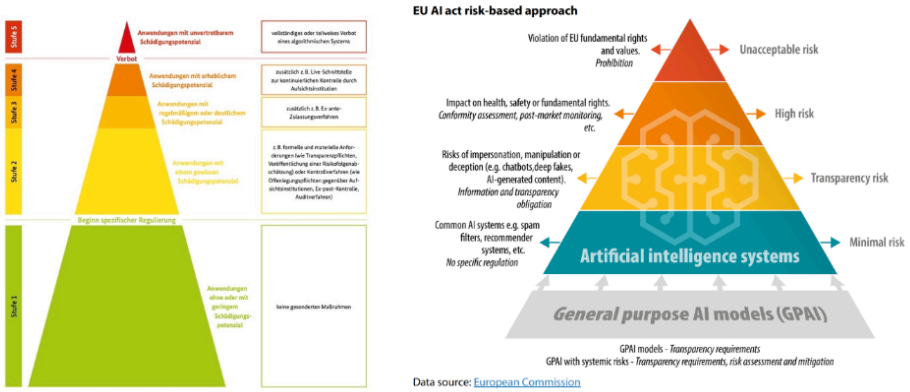


Figure 1 AI risk pyramids of the German Data Ethics Commission and the European Commission/Parliament[42]

The distinction between high-risk and other AI systems is mainly based on the intended purpose,[43] bearing resemblance to other EU product safety regimes that differentiate regulatory requirements according to the assessed risk of a product. For instance, under the Medical Devices Regulation, devices are classified into risk-based cat-

[41] Data Ethics Commission (Germany), Opinion (2020), available at https://www.bfdi.bund.de/SharedDocs/Downloads/EN/Datenschutz/Data-Ethics-Commission_Opinion.pdf, 177.

[42] Sources: left, ibid.; right, European Parliament, Artificial Intelligence Act, Briefing 4th ed., 02-09-2024, https://www.europarl.europa.eu/RegData/etudes/BRIE/2021/698792/EPRS_BRI(2021)698792_EN.pdf.

[43] Almada and Petit (n 2) 94.

egories (i.e., Class I to Class III), with each class subject to increasingly stringent conformity assessments.[44] Similarly, the new Machinery Regulation[45] identifies a list of products subject to higher scrutiny, although it avoids the label "high-risk",[46] ostensibly to prevent the stigmatization of certain products. In contrast, the AIA explicitly introduces the concept of "high-risk" AI systems, but without applying graduated regulation to all lower-risk categories. Instead, AI systems not classified as high-risk are either subject only to transparency obligations or fall outside the scope of binding obligations altogether.

V. The Heuristic Function of the AIA's Risk Pyramid

The visual metaphor of the risk pyramid has remained central in public presentations of the Act. It continues to be used in explanatory materials, policy briefings, and legislative communications to illustrate the AIA's tiered approach to regulating AI systems.[47]

The risk pyramid functions as a policy heuristic—an intuitive visual shorthand that helps policymakers and stakeholders grasp the structure of the AIA at a glance. It communicates the notion that AI systems are categorized by increasing levels of regulatory scrutiny, corresponding to ascending levels of risk. The top of the pyramid represents AI systems considered "unacceptable" due to their potential for serious harm and are, therefore, prohibited. Below this are "high-risk" systems subject to strict requirements, followed by "limited risk" systems for which transparency obligations apply, and finally "minimal risk" systems that remain largely unregulated.

With the introduction of GPAI in Chapter 5 of the AIA, an additional (grey) layer is sometimes appended to the base of the risk pyramid (see Figure 1). Visually, this layer is separated from the more colourful, system-focused tiers above it and connected

[44] Regulation (EU) 2017/745 of the European Parliament and of the Council of 5 April 2017 on Medical Devices amending Directive 2001/83/EC, Regulation (EC) No 178/2002 and Regulation (EC) No 1223/2009 and repealing Council Directives 90/385/EEC and 93/42/EEC [2017] OJ L 117/1. See further Sofia Palmieri, 'The Renewed EU Legal Framework for Medical AI' (2024) 15 European Journal of Law and Technology <https://www.ejlt.org/index.php/ejlt/article/view/969> accessed 31 March 2025; Mathias Karlsen Hauglid and Tobias Mahler, 'Doctor Chatbot: The EU's Regulatory Prescription for Generative Medical AI' (2023) 10 Oslo Law Review 1.

[45] Regulation (EU) 2023/1230 of the European Parliament and of the Council of 14 June 2023 on Machinery and Repealing Directive 2006/42/EC of the European Parliament and of the Council and Council Directive 73/361/EEC, OJ L 165, 29.6.2023, p. 1–102.

[46] The Machinery Regulation refers, in Annex I, instead to "categories of machinery or related products to which one of the procedures referred to in Article 25(2) and (3) shall be applied."

[47] See e.g. European Parliamentary Research Service 2021 report on the AIA: https://epthinktank.eu/2021/11/18/artificial-intelligence-act-eu-legislation-in-progress/artificial-intelligence/

through arrows that imply a functional relationship—namely, that GPAI models may be embedded in or give rise to AI systems positioned elsewhere within the pyramid's regulatory framework. This addition subtly disrupts the logic of the risk pyramid, as it introduces a distinction based not on risk level, but on type—differentiating between AI models and AI systems. Before turning to this structural disruption, however, it is useful to examine more closely the connection between the pyramid and standard risk management approaches.

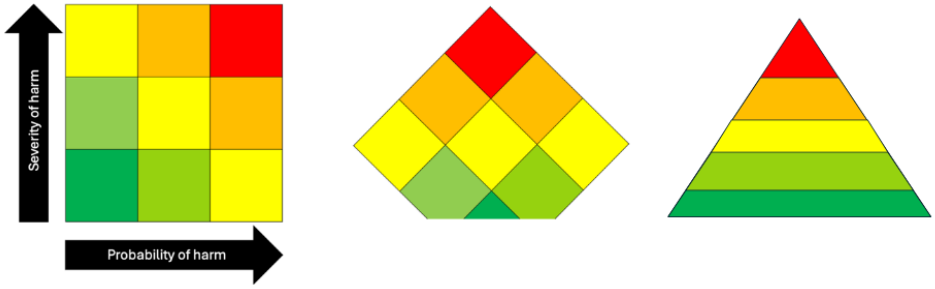


Figure 2 Risk matrix can be transformed into a risk pyramid. Source: Author

The visual structure of the pyramid resonates with tools already familiar in risk management, such as the risk matrix.[48] In that context, risk is typically conceptualized as "the combination of the probability of an occurrence and the severity of the harm," a formulation that appears in Article 3(2) AIA. Risks are often color-coded—red for high, yellow for medium, green for low—and arranged in a matrix that can be turned on its side to form a pyramid-like shape. The graphical simplicity of the pyramid makes it an accessible tool for communicating the AIA's risk-based logic to non-specialists, reinforcing the idea that more perilous AI systems require more intensive regulation.

There is an additional communicative advantage: the risk pyramid effectively serves as a kind of "graphical user interface" for the law. Unlike most legal instruments, which are communicated through dense and often inaccessible text, the pyramid offers a visual schema that purports to explain how the law works.[49] In this respect, it plays a nar-

[48] Mustafa Elmontsri, 'Review of the Strengths and Weaknesses of Risk Matrices' (2014) 4 Journal of Risk Analysis and Crisis Response 49.

[49] On the visualization of law see Arianna Rossi and Monica Palmirani, 'Can Visual Design Provide Legal Transparency? The Challenges for Successful Implementation of Icons for Data Protection' (2020)

rative role, helping to frame the AIA as systematic, proportionate, and modern. It allows lawmakers and regulators to project an image of scientific rationality and regulatory clarity, which is rhetorically powerful—particularly in policy areas characterized by rapid technological development and public unease.

Nevertheless, while the risk pyramid offers an appealing structure for policy communication, one might remain mildly sceptical of how well it captures the AIA's actual legal architecture. This issue will be explored more fully in the following section.

VI. The Risk-Based Narrative

Perhaps surprisingly, the risk pyramid, with all of its various risk levels, is largely absent from the AIA's normative architecture. The legal text explicitly establishes only two risk-based distinctions: first, between high-risk and non-high-risk AI systems; and second, between general-purpose AI models presenting systemic risk and those that do not. The remaining tiers—such as "unacceptable," "limited," and "minimal" risk—are not explicitly codified as formal categories in the legal text. Rather, the risk-based approach operates at the level of policy narrative: it conveys an impression of systematically graduated regulatory responses to varying levels of risk, even though these gradations are not fully embedded in the normative structure of the AIA itself.

The limitations of the risk-based framing become particularly evident in the treatment of prohibited systems. While one might have expected these to be framed as instances of "unacceptable risk," the AIA avoids this terminology. Prohibitions in Article 5 are instead framed categorically. Similarly, the systems subject to transparency obligations under Article 50 are not classified as "limited risk," nor is "minimal risk" a recognised legal term within the AIA. Instead, "non-high-risk" is a relevant category in the AIA, but this is not visible from the visual representation.[50]

The central category of high-risk AI systems is introduced in Chapter III and further specified in Annex III. According to the Commission's impact assessment, this classification was informed by scenarios assessed in terms of risk.[51] However, no formal risk

---

36 Design Issues 82; Rossana Ducato, 'De Iurisprudentia Picturata: Brief Notes on Law and Visualisation Editorial' (2019) 7 Journal of Open Access to Law 1; Tobias Mahler, 'Visualisation of Legal Norms' in Dag Wiese Schartum, Lee Andrew Bygrave and Bekken (eds), Jon Bing: En Hyllest/A Tribute (Gyldendal Norsk Forlag A/S 2014).

[50] Lee A Bygrave and Rebecca Schmidt, 'Regulating Non-High-Risk AI Systems under the EU's Artificial Intelligence Act, with Special Focus on the Role of Soft Law' (Social Science Research Network, 24 October 2024) <https://papers.ssrn.com/abstract=4997886> accessed 15 April 2025.

[51] Impact Assessment to Commission Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act) and Amending Certain Union Legislative Acts, COM (2021) 206 final, (Apr. 21, 2021), footnote 40.

tables or quantified estimates of probability and severity are provided, as would typically be expected in a structured risk assessment. Article 7 does lay out criteria for adding new use cases to Annex III, including contextual and qualitative elements. While these offer relevant considerations, they fall short of constituting a systematic and externally verifiable method for assessing risk level, at least when risk level is expected to be calculated based on "the combination of the probability of an occurrence and the severity of the harm."[52]

The absence of transparent and quantifiable risk reasoning suggests a broader conclusion: the AIA's structure is better understood as the product of political negotiation than as the outcome of a systematic technical or scientific risk evaluation. This suggests that these provisions reflect value-based judgments and deliberative policy choices, rather than risk calculations.[53] In sum, while the pyramid continues to serve as a useful communication tool, it should not be mistaken for a faithful representation of the AIA's legal structure. The classification scheme rests on a conception of AI risk that is only partly transparent, quantifiable, and internally consistent. Rather, it reflects a complex blend of precautionary reasoning, normative priorities, and political compromise—legitimate in their own right, but somewhat removed from what one may expect from a comprehensive and "clearly defined risk-based approach"[54] invoked in the AIA's recitals. Beyond the structural ambiguities discussed so far, new technological developments, particularly the emergence of GPAI, further disrupt the original risk-based framing.

VII. Risk-based GPAI Regulation

The introduction of GPAI models in Chapter 5 AIA further alters the structure implied by the risk-based approach. While the risk pyramid portrays AI *systems* as a one regulatory object to be sorted by risk level—ranging from minimal to unacceptable—GPAI *models* introduce a qualitatively different object of regulation. These models, often referred to elsewhere as "foundation models,"[55] do not serve a specific, predefined

[52] Article 3(2) AIA.

[53] Lilian Edwards, *Regulating AI in Europe: Four Problems and Four Solutions* (Ada Lovelace Institute, March 2022) 11 https://www.adalovelaceinstitute.org/wp-content/uploads/2022/03/Expert-opinion-Lilian-Edwards-Regulating-AI-in-Europe.pdf, p. 11.

[54] Cf. Recital 26 AIA.

[55] David Fernández-Llorca and others, 'An Interdisciplinary Account of the Terminological Choices by EU Policymakers Ahead of the Final Agreement on the AI Act: AI System, General Purpose AI System, Foundation Model, and Generative AI' [2024] Artificial Intelligence and Law <https://doi.org/10.1007/s10506-024-09412-y> accessed 25 April 2025.

function but are instead designed to be adapted for a wide range of downstream applications. This disrupts the logic of risk-based categorization of AI systems based on intended purpose.

GPAI models are not placed within the pyramid's tiered hierarchy but instead form a separate regulatory object. In visual representations of the AIA, they sometimes appear as a grey layer below the pyramid, connected by arrows to indicate that such models may underpin high- or low-risk AI systems.[56] Yet the distinction is not one of risk level but of type: whereas the pyramid classifies systems according to their intended use and associated risks, GPAI regulation is based on model characteristics and the potential for systemic risk.

This distinction is also reflected in the legal architecture. Whereas high-risk AI systems are subject to purpose-specific obligations—such as an obligation to manage risk under Article 9[57]—GPAI models carry purpose-unspecific obligations and a different risk regime for GPAI that presents systemic risk, regulated under Article 55. This article obliges providers to evaluate the model using standardized testing protocols, including adversarial testing, and to identify and mitigate systemic risks that may arise across the lifecycle of the model, similar to the focus of Article 9 for high-risk AI systems.

The distinction between AI systems and GPAI models introduces a dual regulatory track which is not acknowledged by the risk-based approach mainly focusing on high-risk AI systems. Chapter V AIA represents not only a disruption of the visual logic of the pyramid but a departure from the AIA proposal's underlying assumption that AI is a single regulatory object which can be clearly categorized based on risk alone.

VIII. Conclusion: Multiple Approaches to Risk

The risk-based approach adopted in the AI Act is not without merit: it offers a flexible framework intended to tailor regulatory obligations to the level of risk posed by AI systems. However, the coherence suggested by the "risk-based" label dissolves upon closer examination. While the narrative of the risk-based approach is often illustrated through a neatly structured pyramid, the legal reality is more fragmented. In the AIA itself, only two explicit risk-based distinctions are clearly defined—high-risk AI systems and GPAI models representing systemic risk—and these correspond to different domains rather than forming a single, continuous hierarchy. These two categories are disjoint, and the remaining tiers depicted in the pyramid serve primarily a heuristic and communicative function in policy discourse. They are not legally operative in the sense of being tied to clearly defined risk levels or grounded in a standardized method for assessing AI risk.

---

[56] Cf. Figure 1.
[57] Schuett (n 10).

This is also because the idea of defining a risk level once and for all is untenable: risk levels are necessarily constructed differently depending on the actor's perspective, and, in practice, they must primarily be assessed *in concreto*, based on the specific context of use. Indeed, what appears as a single risk-based approach on the AIA's surface is, in fact, a collection of partially overlapping schemes shaped by distinct regulatory objects, where risk is assessed from various actor perspectives. Thus, the Act does not contain one approach to AI risk, but several. This becomes apparent when asking a seemingly simple question: what is the risk level of a given AI system? An AI provider may determine that a system falls into the "high-risk" category under Article 6 because it poses a "significant risk" to health, safety, or fundamental rights, pursuant to Article 6(3). Yet under Article 9, the same system's residual risk must be reduced to a level that is "acceptable"—in practice, likely low or moderate—before it can be placed on the market. At the post-market stage, a market surveillance authority may act under Article 79 or 82 if a system "presents a risk," even where no threshold is defined and the system is otherwise compliant. Further complexity arises from the use of the term "serious risk" in Regulation (EU) 2019/1020, which applies to AI systems as products but is not clearly aligned with the risk categories in the AIA. The result is a risk framework in which the same AI system may be considered "high-risk" (Article 6 AIA), and at the same time be considered of sufficiently low residual risk to be deployed (Article 9(5)), while still potentially representing a "product risk" (Article 79 AIA)—depending on the regulatory context and institutional actor involved. The regulatory conversations[58] about risk within the AIA thus amount to a "risk-regulatory cacophony,"[59] reflecting divergent understandings among lawmakers, enforcement authorities, AI providers, in addition to those affected by AI-driven processes.

Future research should contribute to enhancing conceptual clarity and terminological coherence in the AIA's treatment of risk. It remains an open question whether the risk-based framing advances or undermines the intelligibility of the regulatory structure. Referring to AI systems as "high-risk" risks conflating the act of categorizing AI systems at an abstract level with the concrete assessment of residual risk, and may give the misleading impression of homogeneity across a highly heterogeneous set of technologies. This ambiguity could have been mitigated by adopting a more neutral terminology—such as "regulated AI systems" or "Category A systems"—following the example of the Machinery Regulation, which identifies especially hazardous machinery without employing explicit risk labels.

---

[58] Cf. Julia Black, 'Regulatory Conversations' (2002) 29, Journal of Law and Society, 163-196. https://doi.org/10.1111/1467-6478.00215.

[59] This term refers to the coexistence of multiple, partially inconsistent risk framings across the AIA's regulatory structures.

The analysis of the AIA's risk-based approach highlights several broader insights. It is both notable and commendable that the European lawmaker seeks to apply the concept of risk-based regulation to achieve greater proportionality in the design of legal obligations. However, it must be recognized that a risk-based framing is not inevitable; it remains an optional layer that can be added to regulatory structures to improve their coherence and communicability. Even when only partially implemented, as in the AIA, a risk-based framing can serve a useful heuristic function by offering stakeholders an accessible way of understanding the law's differentiated responses to different types of AI systems. Nevertheless, this strategy is not without cost: the invocation of risk as a central organizing principle may generate conceptual confusion, particularly when the legal text itself does not fully operationalize risk levels through systematic or verifiable criteria. Future legal and policy developments should therefore carefully balance the communicative advantages of risk-based framings against the risks of introducing additional complexity and ambiguity.

Data Governance under the AI Act

Lea Ossmann-Magiera, Lisa Markschies

I. Abstract

This paper examines the data quality requirements established under Art. 10 of Regulation (EU) 2024/1689 (AI Act).[1] Through a detailed analysis of the legal framework in which Art. 10 AI Act is embedded, an assessment of individual data quality and data governance requirements, implementation challenges, and practical considerations, we explore how the AI Act aims to ensure the development of high-quality AI systems. We further analyse to what extent Art. 10 AI Act addresses critical issues such as data bias and contamination. Thereafter, special attention is given to the practical implementation of these requirements through technical standardisation and labelling mechanisms.

II. Art. 10 in the context of the AI Act

Art. 10 AI Act is a cornerstone within the regulation's high-risk obligations framework. To fully understand its significance and scope, it is essential to examine it within the broader context of the AI Act's regulatory structure and its risk-based approach.

1. The AI Act's risk-based approach

The AI Act employs a risk-based approach to regulation, where legal obligations are determined according to the level of risk posed by an AI system. While the risk classification has been carried out in advance by the legislator, it is up to the providers of AI systems to assess in each individual case which of the risk classes their system is to be categorised in. The regulation establishes four risk categories:

[1] Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) No 300/2008, (EU) No 167/2013, (EU) No 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Artificial Intelligence Act).

1. Unacceptable Risk: AI systems in this category are prohibited within the EU (Art. 5 AI Act).
2. High Risk: AI systems are subject to specific diligence obligations (Art. 8-15 AI Act).
3. Limited Risk: AI systems are subject to transparency obligations (Art. 50 AI Act).
4. Minimal Risk: AI systems are governed through self-regulation (Art. 95 AI Act).

Art. 10 AI Act thus falls within the high-risk category as part of the comprehensive diligence obligations framework, alongside requirements for risk management, technical documentation, record-keeping, transparency, human oversight, and IT security. Notably, this risk-based framework applies exclusively to AI systems and not to General Purpose AI (GPAI) models, which are subject to separate provisions.[2]

All obligations listed in Art. 8-15 AI Act apply to the entire life cycle of high-risk AI systems. They, hence, aim to capture the development and training phase before placing the AI system on the market as well as the phase after placement on the market as part of the market surveillance obligations. While Art. 10 AI Act primarily focuses on the development phase — covering data collection, preprocessing, training, testing, evaluation, and software integration — it extends to deployment scenarios where training, validation, or testing occurs. This comprehensive approach reflects the legislator's recognition that data quality must be maintained across all stages of an AI system's existence.

2. Value chain complexity and regulatory challenges

Today's AI value chains present unique regulatory challenges due to their length and complexity. Such a value chain typically encompasses the model development phase, model deployment phase, AI system development, and AI system deployment phase. It can be visualised as follows:

A critical regulatory challenge arises from the fact that the entity developing an AI model may differ from the one placing the model or the later AI system on the market. The AI Act addresses this through a layered approach to obligations. First, GPAI model providers must adhere to data quality obligations under Art. 53 AI Act in conjunction with Annex XI 2(c) and Annex XII 2(c) AI Act. Second, downstream providers who

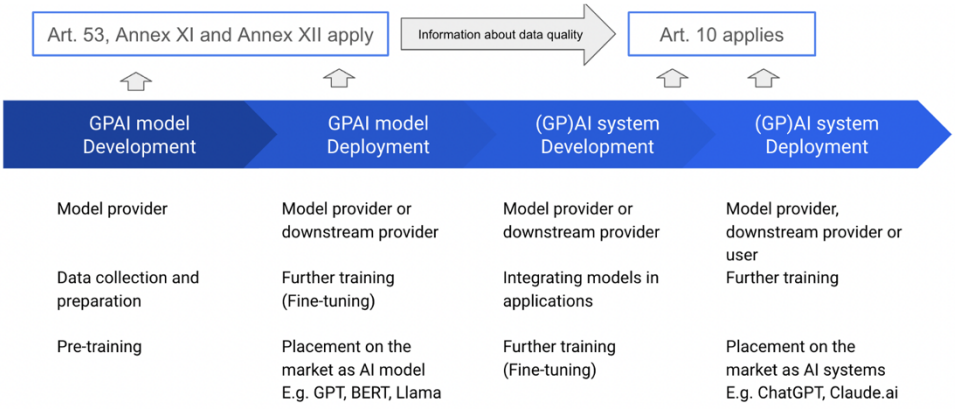
[2] See Art. 53-55 AI Act.

fine-tune models for specific purposes are subject to Art. 10 requirements. Third, entities that either place their name/trademark on a high-risk AI system, make substantial modifications to existing systems, or deploy systems with continued training capabilities must also comply with Art. 10's provisions.

This comprehensive regulatory approach along the value chain is intended to address the "garbage in, garbage out" problem, acknowledging that poor data quality during development inevitably leads to compromised output quality. Poor output quality can, in turn, lead to both personal injury and property damage. The regulation attempts to maintain data quality standards by imposing obligations on all actors who can exert meaningful influence over the system's data handling practices.

Although the focus of this paper is on Art. 10 AI Act — i.e. obligations for providers of high-risk systems — it is essential to take into account the entire value chain for meaningful data quality. It will become clear in the coming months and years to what extent the information that providers of GPAI models have to pass on at the beginning of the chain according to Art. 53 AI Act is sufficient. Art. 53 AI Act was, in fact, included at the last minute of the legislative process,

At the time of writing this article, the AI Office has initiated the process of drawing up Codes of Practice (CoP). Such CoP — which are elaborated in a multi-stakeholder



consultation — are aimed at specifying the requirements set out in Art. 53 AI Act and are scheduled for completion in the first half of 2025.

III. Data governance and data quality obligations in depth

The following section provides a detailed examination of each paragraph within Art. 10 AI Act, analysing the specific obligations and requirements imposed on AI system

providers. The provisions establish a framework for ensuring data quality through various mechanisms, including data governance, documentation requirements, and quality control measures.

1. Scope and Fundamental Provisions

Art. 10's first paragraph establishes critical limitations and fundamental principles for data quality in high-risk AI systems. Art. 10 (1) AI Act stipulates that the data quality requirements of Art. 10 AI Act exclusively apply to high-risk AI systems and specifically targets training, validation, and testing data. It, hence, focuses on data-based AI techniques involving training (i.e. machine learning techniques). With regard to logic- and knowledge-based approaches not relying on training datasets, Art. 10 AI Act only refers to the quality of testing data sets (see Art. 10 (6) AI Act). Input data, on the other hand, is not addressed by the data quality and governance requirements set out in Art. 10 AI Act.

Furthermore, Art. 10 (1) AI Act formulates the general rule that datasets must meet specific quality criteria. The paragraph further mandates providers to adhere to data governance and management practices specified in paragraph 2, to quality criteria detailed in paragraphs 3 and 4, and to establish governance practices elaborated in paragraph 2. Furthermore, it introduces an exception to Art. 9 of the Regulation (EU) 2016/679 (GDPR)[3] regarding data processing.

2. Data governance and management practices

Art. 10 (2) AI Act provides a non-exhaustive list of data governance and management practices.[4] These include choosing the relevant design, taking into account data collection processes and the origin of data, preparing data processing operations appropriately,[5] formulating assumptions with respect to the information that the data are supposed to measure and represent, assessing the availability, quantity, and suitability of the data sets that are needed, as well as practices that aim at identifying, preventing and mitigating possible biases.[6]

A thorough data management in the beginning of the AI system's life cycle — in the development stage — is essential, since the AI system's output quality will only be as high as the quality of the data it is trained with. Hence, the first step captured by Art.

[3] Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation).

[4] Müller-Peltzer and Tanczik, Künstliche Intelligenz und Daten, RDi 2023, 452, 455; Braun Binder and Egli, in Martini and Wendehorst, KI-VO, 1st ed. 2024, Art. 10 mn. 16.

[5] E.g. through annotation, labelling, cleaning, updating, enrichment, and aggregation.

[6] Biased data will be discussed in detail in section IV.1.

10 (2) AI Act is the preparation of training data sets. Training data sets are usually created and curated by natural persons.[7] When collecting, preparing, and processing these data sets, numerous decisions are to be made. These decisions have consequences for the functionality and accuracy of the data sets. If the initial training data sets are not carefully selected, they can transmit and even perpetuate and increase biases or discrimination towards vulnerable groups that are present in our society.

Validation and testing data sets must also be subject to data governance and management practices outlined above in order to guarantee for an efficient evaluation of data quality before the AI system may be placed on the market.

The practices that Art. 10 (2) AI Act refers to must be *appropriate* for the intended purpose. What *appropriate* means in this context is not specified in the AI Act.[8] Art. 10 (2) AI Act merely provides a non-exhaustive list of working steps where data governance and management practices should be applied.[9] These start at the very early stage of choosing the model's design and continue throughout the entire life cycle as long as the system is being trained, validated, or tested. The practices themselves and what exactly they should entail is not explained in the recitals. Art. 10 (2) AI Act is therefore one of the provisions of the AI Act that is subject to undefined legal terms and, therefore, requires interpretation. Different approaches to specify such undefined legal terms in practice is discussed below.

The only specification made in the recitals stipulates that the requirements related to data governance can be complied with by having recourse to third parties that offer certified compliance services including verification of data governance, data set integrity, and data training, validation, and testing practices.

3. Data quality requirements

According to Art. 10 (3) AI Act training, validation, and testing data sets must be relevant, sufficiently representative, and to the best extent possible, free of errors and complete in view of the intended purpose. Paragraph 3, hence, articulates concrete data quality requirements that blend traditional and AI-specific considerations. While "classic" data quality requirements encompass relevant, error free, and complete data, representativeness can be considered an "AI-specific" data quality requirement.[10] These data quality requirements do not have to be fulfilled completely, in the sense that data must be, e.g., completely free of errors. Instead, they are to be fulfilled gradually, which

[7] Müller-Peltzer and Tanczik, Künstliche Intelligenz und Daten, RDi 2023, 452, 456.

[8] Braun Binder and Egli, in Martini and Wendehorst, KI-VO, 1st ed. 2024, Art. 10 mn. 15.

[9] Müller-Peltzer and Tanczik, Künstliche Intelligenz und Daten, RDi 2023, 452, 455; Braun Binder and Egli, in Martini and Wendehorst, KI-VO, 1st ed. 2024, Art. 10 mn. 16.

[10] Braun Binder and Egli, in Martini and Wendehorst, KI-VO, 1st ed. 2024, Art. 10 mn. 65.

means that the requirements can be fulfilled to different extents. Such a gradual approach to data quality is common practice in the fields other than AI.

The concept of data relevance is a critical criterion, measuring data's applicability and utility for specific tasks. Relevance hence means the degree to which data is applicable and useful for the respective task. An AI system designed for diagnosing skin cancer, for example, should be trained on dermatological images from a wide variety of skin tones and types. Irrelevant data, such as images of non-skin-related conditions, would harm the performance and accuracy of the AI system. This requirement is particularly challenging in unsupervised machine learning contexts, where the technology's exploratory nature often means that relevant correlations are not immediately apparent.

Representativeness stands as another key requirement, ensuring that datasets adequately represent the operational context and population. For example, an AI system designed for diagnosing skin cancer should be trained on dermatological images from a wide variety of skin tones and types. The training data must represent all skin types and not over or underrepresent certain types. This provision aims to prevent systemic biases by demanding balanced data representation. The regulation pragmatically recognises the impossibility of absolute representativeness, instead requiring "sufficiently" representative data.

Accuracy and completeness are further refined through a nuanced approach. Accuracy means that the information contained in the data corresponds to reality. A data set is complete if values are available for each attribute for all recorded entities.[11] For example, the training data for all persons (entities) must include their year of birth (attribute) if this is to be relevant. The requirement that data correspond to reality “to the extent possible" acknowledges the inherent limitations of error-free data in information technology. Similarly, the completeness requirement allows for practical limitations, recognising that not every dataset can and must be perfectly comprehensive. This amendment, which was added in the course of the regulatory process, is very important because it is not feasible to guarantee for completely error-free data.[12] In information technology there are no error-free models, only more or less precise mathematical models.[13] In addition, society's tolerance for errors also varies depending on the area of application of the high-risk AI system.[14] Furthermore, errors often only appear in the validation phase (especially with unsupervised learning) and guaranteeing for completely error-free training data sets is not practicable.

---

[11] Braun Binder and Egli, in Martini and Wendehorst, KI-VO, 1st ed. 2024, Art. 10 mn. 72.

[12] Bomhard and Merkle, Europäische KI-Verordnung, RDi 2021, 276, 280; Braun Binder and Egli, in Martini and Wendehorst, KI-VO, 1st ed. 2024, Art. 10 mn. 69; Spindler, Anforderungen an Hochrisiko-KI-Systeme, in Hilgendorf and Roth-Isigkeit, Die neue Verordnung der EU zur Künstlichen Intelligenz, 1st ed. 2023, § 5 mn. 29.

[13] Quelle Informatik

[14] Braun Binder and Egli, in Martini and Wendehorst, KI-VO, 1st ed. 2024, Art. 10mn. 69.

A practical aspect is the provision that allows collective dataset performance, rather than demanding perfection from individual datasets. This approach recognises the iterative nature of AI system development and the limitations of early-stage data understanding.

Contextual considerations in paragraph 4 further refine these requirements. Providers must train systems specific to intended market contexts, with considerations limited to typical use settings. The provider must therefore not rely on the fact that his system will also function flawlessly in a different environment but must train it specifically for the respective regional, linguistic, cultural, or objective context of that market.[15] This approach prevents providers from claiming universal applicability while excluding atypical circumstances and foreseeable misuse. Paragraph 4 therefore specifies the requirement of data relevance.[16]

4. Exception from the processing of special categories of personal data

The fifth paragraph of Art. 10 AI Act addresses the sensitive issue of special category data processing. According to Art. 9 GDPR, it is prohibited to process particularly sensitive categories of personal data, such as health data, data about sexual orientation and others. While maintaining the general prohibition on processing sensitive personal data, the AI Act provides a specific exception for bias detection and correction in Art. 10 (5) AI Act.

As a general rule, the AI Act does not affect the provisions of the GDPR (which leads to a parallel application of both legal frameworks).[17] Furthermore, the AI Act does not constitute a legal basis for processing personal data in the sense of Art. 6 GDPR.[18] Consequently, another legal basis such as consent or a legitimate interest is required if personal data is processed during the training of AI systems. Art. 10 (5) AI Act does so by explicitly stipulating a specific exception for the processing of special categories of personal data for reasons of bias detection and correction. However, this exception only applies if the data processing is strictly necessary to detect and correct biases. No other purpose can justify the processing of sensitive personal data.[19]

---

[15] Braun Binder and Egli, in Martini and Wendehorst, KI-VO, 1st ed. 2024, Art. 10 mn. 77; Grützmacher CR 2021, 433 (440)

[16] Spindler, Anforderungen an Hochrisiko-KI-Systeme, in Hilgendorf and Roth-Isigkeit, Die neue Verordnung der EU zur Künstlichen Intelligenz, 1st ed. 2023, § 5 mn. 29; Braun Binder and Egli, in Martini and Wendehorst, KI-VO, 1st ed. 2024, Art. 10 mn. 77.

[17] See Art. 2 (7) AI Act.

[18] See recital 63 AI Act.

[19] Spindler, Anforderungen an Hochrisiko-KI-Systeme, in Hilgendorf and Roth-Isigkeit, Die neue Verordnung der EU zur Künstlichen Intelligenz, 1st ed. 2023, § 5 mn. 32.

In addition to that, providers have to guarantee for appropriate safeguards which are listed in Art. 10 (5) (a)-(f) AI Act. The list has been added during the legislative process.

IV. Current challenges

The importance of high-quality data for the development of AI systems is highlighted when looking at specific problems that arise due to insufficient data. The following section discusses two current issues that may occur in this context: biased AI (IV. 1.) and model collapse due to contaminated data (IV. 2.); it then looks at the possible solutions offered by Art. 10 AI Act.

1. Biased AI

There is no doubt that the real world is heavily biased and while there may have once been a perception that machines do not display those same very human traits, reality paints a different picture. It has been demonstrated that algorithms can discriminate against women when evaluating job applications[20] or fail to diagnose skin cancer in non-white patients.[21] A general understanding has settled in that decisions made by AI systems may not be the fairest after all.[22] One of the main causes of biased AI appears to be biased data,[23] which raises the question of how this phenomenon can be counteracted. Art. 10 AI Act seems to promise regulatory solutions as it creates data governance and quality requirements.[24] However, the effectiveness of these measures may be less than initially anticipated.

[20] Dastin, Insight - Amazon scraps secret AI recruiting tool that showed bias against women, Reuters, 11. October 2018, https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G/.

[21] Daneshjou et al., Disparities in dermatology AI performance on a diverse, curated clinical image set, Science Advances 2022, Vol. 8 (32); Wen et al., Characteristics of publicly available skin cancer image datasets: a systematic review, Lancet Digital Health 2022, Vol. 4 (1), e64-e74.

[22] Kern, Humans versus machines: Who is perceived to decide fairer? Experimental evidence on attitudes toward automated decision-making, Patterns 2022, Vol. 3 (10); Hajigholam Saryazdi, Algorithm Bias and Perceived Fairness: A Comprehensive Scoping Review, in Zaza et al. (eds.), SIGMIS-CPR '24: Proceedings of the 2024 Computers and People Research Conference, 2024, Article No. 6, p. 1-9.

[23] For an overview of causes and mitigation of biased AI see Pagano et al., Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods, Big Data and Cognitive Computing 2023, Vol. 7 (1).

[24] See above Sec. III.

a) *What are biases and how are they typically addressed*

Firstly, it is essential to gain an understanding of what can be understood by "bias". The AI Act specifically mentions bias in a number of instances.[25] However, despite those references, the AI Act fails to offer a real definition. For the purpose of this paper, bias in the context of AI can be generally understood as systemic errors that favour certain groups while disadvantaging others.[26] The occurrence of these biases can be attributed to a number of factors. One example is the utilisation of data that has already been a subject of biases, which is then employed in the development of AI. In the aforementioned example of reactions to job applications, it can be observed that the majority of previous hires were male. Such data may be interpreted by an AI system as evidence that gender is a determining factor in hiring decisions, perpetuating historical biases and contributing to the persistence of existing inequalities. Although biased data is not the sole cause of biased AI, it can have a significant impact on the performance of an AI system at an early stage of the value chain.[27]

Some biases may manifest in a way that constitutes a case of direct or indirect discrimination, which may resemble those previously observed in the analogue world like the rejection of job applications based on gender and which are therefore evident to a certain degree. In other cases, the existence of bias and the resulting consequences may be more difficult to discern. In particular, generative AI could produce content where bias is less overt, for instance, by inadequately representing groups of people due to unbalanced content and non-inclusive language.[28]

In many instances of discrimination by biased AI systems, anti-discrimination legislation may be applicable.[29] However, there are challenges in enforcing "analogue" anti-discrimination laws in the context of AI systems. This is because such laws are typically designed to address individual cases that arise after the deployment of an AI system.[30] They are therefore not tailored to the characteristics of algorithmic discrimination. The

[25] In addition to Art. 10 AI Act, the term "bias" is referenced in Art. 14 (4) (b), Art. 15 (4), Annex XI (1) (2) (c) AI Act and various recitals.

[26] Bellamy et al., AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias, arXiv 2018, arXiv:1810.01943.

[27] Pagano et al., Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods, Big Data and Cognitive Computing 2023, Vol. 7 (1).

[28] Hacker et al., Generative Discrimination: What Happens When Generative AI Exhibits Bias, and What Can Be Done About It, forthcoming in Hacker, Engel, Hammer, and Mittelstadt (eds.), The Oxford Handbook of the Foundation and Regulation of Generative AI, 2025.

[29] Von Ungern-Sternberg, Discriminatory AI and the Law – Legal Standards for Algorithmic Profiling, in Voeneky, Kellmeyer, Mueller, and Burgard (eds.), The Cambridge Handbook of Responsible Artificial Intelligence, 2022, p. 252.

[30] Deck et al., Implications of the AI Act for Non-Discrimination Law and Algorithmic Fairness, arXiv 2024, arXiv:2403.20089, p. 2.

AI Act might bridge this enforcement gap, as it doesn't look at discrimination from an ex post perspective but rather demands fairness to be taken into account at the earliest stages of model development, thus potentially preventing biases to occur in the first place.[31]

b) *Measures taken in Art. 10 AI Act*

The increased awareness of the risks associated with biased AI and the willingness to establish obligations with the aim of mitigating those risks before they manifest can already be considered an advancement from the status quo. Art. 10 AI Act lays out criteria meant to ensure that the data used in AI development is less biased and as a result creates less biased AI. It remains to be seen whether the measures taken by the European legislator will fulfil those promises. The following chapter will examine the opportunities and potential pitfalls.

The first provision to explicitly tackle biased AI is Art. 10 (2) AI Act. Herein, the obligation for data governance and management practices to examine "possible biases that are likely to affect the health and safety of persons, have a negative impact on fundamental rights or lead to discrimination prohibited under Union law, especially where data outputs influence inputs for future operations" is established. The amendments to this wording throughout the legislative process serve to illustrate that the primary objective of this provision was the prevention of discrimination.[32] Building on this, Art. 10 (2) (g) AI Act then stipulates the necessity for demonstrated measures related to the detection, prevention, and mitigation of those discovered possible biases to be taken. As shown above, the obligations of Art. 10 (2) AI Act are rather vague, and the corresponding recitals do not provide further clarifications.[33] What is or isn't an *appropriate* measure might vary greatly depending on the specific circumstances of the case. Consequently, the obligations set out in Art. 10 (2) (f) and (g) AI Act may prove challenging to navigate, given the surrounding legal uncertainty.

Art. 10 (5) AI Act may provide a means of facilitating compliance with Art.10 (2) (f) and (g) AI Act as it allows for the processing of special categories of personal data possible in order to debias AI systems. In terms of the efficacy of this stipulation, one might inquire as to why the decision was taken to restrict this exception to the special categories of personal data pursuant to Art. 9 (1) GDPR, instead of including all kinds of personal data. While AI may inhibit biases that correlate with the attributes listed in Art. 9 (1) GDPR, like racial origin, sexual orientation, or health, others that are not included in this list, for example age or gender, may be just as significant. It is possible

[31] "Enforcement by design" Deck et al., Implications of the AI Act for Non-Discrimination Law and Algorithmic Fairness, arXiv 2024, arXiv:2403.20089, p. 2.

[32] Braun Binder and Egli, in Martini and Wendehorst, KI-VO, 1st ed. 2024, Art. 10 mn. 27.

[33] See above Sec III. 2.

that the legislator assumed that the processing of "regular" personal data would be covered by one of the legal grounds set out in Art. 6 GDPR and therefore didn't need to be included in Art. 10 (5) AI Act. Regarding legal certainty and the effectiveness of Art. 10 (5) AI Act, it would have been beneficial to establish legal grounds for the processing of both categories of personal data for the purpose of debiasing. Nevertheless, even if it is assumed that the processing of regular personal data will be generally easily possible, the question whether Art. 10 (5) AI Act is a suitable tool for debiasing AI remains. One aspect that raises doubts about this is the fact that the hurdles that have to be taken in order to invoke this exception are relatively high.[34] The processing of special categories of data may only occur "to the extent that it is strictly necessary for the purpose of ensuring bias detection and correction." The strict necessity is a condition that may prove challenging to fulfil. Although the addition of the catalogue of safeguards pursuant to Art. 10 (5) (a) to (f) AI Act throughout the legislative process was essential to guarantee adequate data protection, it might be the reason that the opportunity to conduct exceptional processing of those special categories of data will rarely be exercised. In particular, the stipulation that the bias detection and correction cannot be effectively fulfilled by processing other data, including synthetic or anonymised data (Art. 10 (5) (a) AI Act), will, once more, most likely be difficult to fulfill — especially before such processing has taken place. Providers may be disinclined to make use of this exception as they have to fear that even though they might assume that the processing is strictly necessary and cannot be achieved by processing other kinds of data, this assumption might turn out to be wrong, making the processing unlawful.

In conclusion, Art. 10 AI Act may serve as a foundation for the utilisation of less biased data in the development of AI, which could ultimately result in the creation of AI systems that are less biased. The practical effectiveness of the approaches chosen might be limited by legal uncertainty and strict requirements for the processing of special categories of data. Data governance is of significant importance in the mitigation of biases. However, technical advancements in debiasing AI that offer solutions that are not necessarily dependent on unbiased data sets shouldn't be disregarded.[35] It is also important to note that non-high-risk AI, which typically includes generative AI, is not covered by Art. 10 AI Act and is therefore not subject to the same standards when it

[34] In depth: Surjadi, Die Rechtmäßigkeit der Verarbeitung sensibler Daten nach Art. 10 Abs. 5 AI Act – Ein Durchbruch für das Debiasing von KI-Systemen?, in Heinze and Steinrötter (eds.), KI und Daten: Digitalregulierung auf dem Höhepunkt?, 2024.

[35] Pagano et al., Bias and Unfairness in Machine Learning Models: A Systematic Review on Datasets, Tools, Fairness Metrics, and Identification and Mitigation Methods, Big Data and Cognitive Computing 2023, Vol. 7 (1).

comes to the detection and mitigation of biases.[36] This may prompt the question whether those types of biases can be seen as not needing the same amount of regulation. Nevertheless, the general shift towards an "enforcement by design"[37] is a positive development.

2. Model collapse due to contaminated data

A further potential challenge at the intersection of data governance and AI development is that of contaminated data, which has only been the subject of discussion for a relatively short period of time.

a) *What is contaminated data and why is it a problem?*

The phenomenon of model collapse due to contaminated data was notably described by Shumailov et al. in mid-2024.[38] It was demonstrated that training models on synthetic data could result in models generating nonsensical or simply less desirable output, indicating early stages of model collapse.[39] This is especially problematic as there is now a vast amount of AI-generated content that floods the internet and gets ingested when techniques such as web scraping are used in order to accumulate training data. This also implies that the current market leaders, who were among the first to adopt this technology, have a competitive advantage due to their access to human-generated content from before the widespread use of generative AI, while simultaneously causing the landscape to be overtaken by artificial content. Those attempting to enter the market today are confronted with a markedly disparate online environment. Furthermore, data contamination differs from data poisoning attacks. In contrast to data poisoning attacks, which are typically targeted and driven by malicious intent,[40] data contamination represents a symptom of a broader shift in the online environment. Consequently, it is not really a cybersecurity issue but rather a more general phenomenon. It is for this reason that data poisoning is addressed in Art. 15 (5) AI Act which focuses on robustness.

---

[36] Hacker et al., Generative Discrimination: What Happens When Generative AI Exhibits Bias, and What Can Be Done About It, forthcoming in Hacker, Engel, Hammer, and Mittelstadt (eds.), The Oxford Handbook of the Foundation and Regulation of Generative AI, 2025.

[37] Deck et al., Implications of the AI Act for Non-Discrimination Law and Algorithmic Fairness, arXiv 2024, arXiv:2403.20089.

[38] Shumailov et al., AI models collapse when trained on recursively generated data, Nature 631 (2024), 755-759.

[39] The extent of this has been questioned by other scholars, see Gerstgrasser et al., Is Model Collapse Inevitable? Breaking the Curse of Recursion by Accumulating Real and Synthetic Data, arXiv 2024, arXiv:2404.01413 and Kazdab et al., Collapse or Thrive? Perils and Promises of Synthetic Data in a Self-Generating World, arXiv 2024, arXiv:2410.16713.

[40] Schwarzschild et al., Just How Toxic is Data Poisoning? A Unified Benchmark for Backdoor and Data Poisoning Attacks, in Meila and Zhang (eds.), Proceedings of the 38th International Conference on Machine Learning, 2021, Vol. 139, p.9389-9398.

Therefore it is evident that the measures implemented to mitigate the risks of data poisoning cannot be directly transferred to data contamination.

b) *Remedies within the AI Act*

In seeking solutions to this issue within the AI Act, Art. 10 AI Act could be the starting point. However, the scope of this article may prove inadequate for effectively addressing the issue of contaminated data.[41] The provision is aimed at providers of high-risk AI systems[42] which presents a significant challenge. Firstly, the majority of AI systems will not fall under the high-risk categorisation,[43] leaving, for example, LLMs that utilise extensive web scraping uncovered by this provision. Secondly, the obligations are primarily directed towards the AI provider. It does not create direct due diligence obligations for data providers,[44] even less for competitors who "pollute" the online landscape by generating artificial output. Instead, emerging new competitors offering high-risk AI systems may even be prohibited to enter the market if they cannot rid their data sets of the contaminated data, and if artificial data can be seen as constituting an error pursuant to Art. 10 (3) AI Act. Art. 10 AI Act might therefore not be a solution after all but rather perpetuate existing power imbalances. The rules for general purpose models, laid down in Art. 51 et seq. AI Act are also of little help as those only establish documentation obligations with regards to the data used in training, Art. 53 (1) (a), Annex XI (Sec. 1) (2) (c) AI Act. Another provision of the AI Act that might prove useful is Art. 50 (2) AI Act, which obligates the providers of AI systems that generate synthetic audio, image, video, or text content to ensure that the outputs are marked as artificially created, making it easier to identify and avoid AI-generated output when accumulating training data. This makes it easier to identify and avoid AI-generated output when accumulating training data. However, this does not address the issue of content that is already in circulation, nor is it applicable to content generated outside of the EU. In conclusion, it can be stated that the AI Act and, in particular, Art. 10 AI Act do not

[41] Burden et al., Legal Aspects of Access to Human-Generated Data and Other Essential Inputs for AI Training, University of Cambridge Faculty of Law Research Paper, No. 35/2024, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5045155.

[42] See above Sec III. 1.

[43] For estimations on the percentage of AI systems that will be classified to be high-risk, see appliedAI initiative, AI Act: Risk Classification of AI Systems from a Practical Perspective, 2023, https://aai.frb.io/assets/files/AI-Act-Risk-Classification-Study-appliedAI-March-2023.pdf.

[44] Those obligations can be made part of a contract between the provider of a system and the data supplier. In this case general contract and tort laws apply. However a causal link between contaminated data and damages needs to be established for liability rules to apply, cf. Burden et al., Legal Aspects of Access to Human-Generated Data and Other Essential Inputs for AI Training, University of Cambridge Faculty of Law Research Paper, No. 35/2024, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5045155.

provide the necessary tools to address the issue of potential model collapse due to the presence of contaminated data.

V. Practical implementation

The last section of this paper focuses on the question of how the requirements set out in Art. 10 AI Act are to be implemented in practice.

1. Standardisation

Technical standardisation plays a crucial role in operationalising the legal requirements of Art. 10 AI Act. Building on well-established practices in product safety law, standardisation provides concrete methods for demonstrating compliance with legal requirements. This approach offers practical guidance while maintaining flexibility for technological advancement.

The legal foundation for technical standardisation in the EU is established by Regulation 1025/2012[45]. Standardisation offers a primary advantage through the mechanism of presumption of conformity. According to Art. 40 AI Act, conformity with the AI Act's provisions is presumed if providers adhere to harmonised European Norms (hEN). Such hEN serve to clarify the often ambiguous and indefinite legal provisions inherent in complex regulatory frameworks, which is especially significant for technology providers seeking clear guidance.

In December 2022, a standardisation request pertaining to Art. 8-15 AI Act was formally issued by the EU Commission, initiating a critical process of developing comprehensive technical standards. However, the current status remains uncertain, with ongoing work on the standards raising questions about their potential timely completion. The developing standardisation process represents a complex intersection of legal requirements and technological innovation, where the ultimate effectiveness depends on the precise articulation of technical specifications and compliance mechanisms.

2. Labelling

Another tool to facilitate easier compliance with Art. 10 AI Act is the establishment of data quality and utility labels. This kind of label has been introduced in Art. 56 of the Proposal for a European Health Data Space (EHDS) and explicitly mentions Art. 10

[45] Regulation (EU) No 1025/2012 of the European Parliament and of the Council of 25 October 2012 on European standardisation, amending Council Directives 89/686/EEC and 93/15/EEC and Directives 94/9/EC, 94/25/EC, 95/16/EC, 97/23/EC, 98/34/EC, 2004/22/EC, 2007/23/EC, 2009/23/EC and 2009/105/EC of the European Parliament and of the Council and repealing Council Decision 87/95/EEC and Decision No 1673/2006/EC of the European Parliament and of the Council.

AI Act. The development of AI is one of the secondary uses for which health data may be obtained by national health data access bodies. AI systems that are used in the medical sector will also generally fall into the high-risk category and thus be subject to the obligations of Art. 10 AI Act. For those reasons the introduction of a data quality label appears to be a logical step, creating synergies between both the AI Act and the EHDS. In the future, the European legislator has the opportunity to create such data labels for other kinds of data, particularly by making use of the envisaged data spaces. Greater certainty regarding the quality of data can facilitate innovation.

VI. Conclusion

This paper has examined the data quality requirements established under Art. 10 AI Act. First, it provided an overview of the AI Act's regulatory structure in which Art. 10 AI Act is embedded. After that, an assessment of individual data quality and data governance requirements, implementation challenges, and practical considerations were explored. It was further analysed to what extent Art. 10 AI Act addresses critical issues such as data bias and contamination. Lastly, special attention was given to the practical implementation of these requirements through technical standardisation and labelling mechanisms.

Art. 10 AI Act is therefore an essential provision of the AI Act that addresses the quality of AI output. The extent to which the rule will be able to solve pressing problems such as biased data and contamination remains to be seen. However, it is to be welcomed that the AI Act at least touches on these issues.

Human Oversight under the AI Act and its interplay with Art. 22 GDPR

Tristan Radtke

The AI Act identifies human oversight as a fundamental mechanism for human control over AI systems to protect health, safety, and fundamental rights. While this may be promising from a policy perspective, there are some challenges to achieving effective human oversight. This Article sheds some light on how the concept of human oversight might fit into the AI Act's overall regulatory approach to human-centric AI. Based on the concept of human oversight outlined by the AI High-Level Research Group, the Article analyzes the classification of human oversight measures under the AI Act, as well as the obligations and challenges arising from these provisions. In this context, particular attention is paid to the right under Art. 22 GDPR. Despite the challenges of the concept of human oversight, the different approaches under the AI Act and Art. 22 GDPR may work well together. Nevertheless, the effectiveness of human oversight depends in particular on effective tools in practice, and it remains to be seen whether the solution here lies in technology.

The concept of human oversight needs to be approached from a policy perspective first (see below, sub. I.), before analyzing the specific obligations for providers and deployers of AI systems (see below, sub. II.), the interplay of these provisions with other legal acts (see below, sub. III.) and potential challenges for effective human oversight (see below, sub. IV.).

I. Policy reasoning – Human oversight as a principle in the AI Act and the GDPR?

First, the policy underlying the AI Act's concept of human oversight will be addressed. This requires taking into account the history of the AI Act, previous statements by relevant stakeholders and Art. 22 GDPR.

1. Ethics Guidelines by the High-Level Expert Group on Artificial

Prior to the drafting and adoption process of the AI Act, the European Commission appointed a group of experts to advise on the Commission's AI strategy, the AI High-Level Expert Group (AI HLEG).[1] One of the most important deliverables for trustworthy AI and the necessary principles is the Ethics Guidelines for Trustworthy AI.[2] In these guidelines, the AI HLEG identifies seven non-exhaustive aspects that it believes are essential for trustworthy AI based on the fundamental rights framework[3] and ethical principles[4]. As the first such aspect, the AI HLEG names "human agency and oversight including fundamental rights, human agency and human oversight".[5] The AI HLEG sees human oversight as a concept for tools to ensure that AI stays within the legal, ethical, and societal boundaries, i.e. that it "does not undermine human autonomy or causes other adverse effects"[6] and that human agency is maintained.[7]

2. AI Act

The AI Act[8] explicitly refers to the AI HLEG and the seven non-binding aspects, including human oversight, in Recital 27. While all aspects are listed in Recital 27 (3) AI Act, the legislator particularly emphasizes human oversight in Recital 27 (5) AI Act and considers the concept important in order to use AI "as a tool that serves people, respects human dignity and personal autonomy, and that is functioning in a way that can be appropriately controlled and overseen by humans".

The AI Act establishes "human-centric [...] AI" (Art. 1 (1) AI Act) as a central goal of the Act.[9] This requires, one the hand, making the best possible use of the potential of AI systems, and, on the other hand, limiting the autonomy of AI systems in order to

[1] See https://digital-strategy.ec.europa.eu/en/policies/expert-group-ai. All internet references were last accessed on 8 January 2025. On the AI HLEG Guidelines Dettling/Krüger, Erste Schritte im Recht der Künstlichen Intelligenz – Entwurf der „Ethik-Leitlinien für eine vertrauenswürdige KI", MMR 2019, 211.

[2] HLEG-AI, Ethics Guidelines for Trustworthy AI, 2019, https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai.

[3] HLEG-AI, Ethics Guidelines for Trustworthy AI, 2019, p. 10-11: including human dignity, freedom of the individual, democracy, equality and citizens' rights.

[4] HLEG-AI, Ethics Guidelines for Trustworthy AI, 2019, p. 11.

[5] HLEG-AI, Ethics Guidelines for Trustworthy AI, 2019, p. 14.

[6] HLEG-AI, Ethics Guidelines for Trustworthy AI, 2019, p. 16.

[7] Cf. HLEG-AI, Ethics Guidelines for Trustworthy AI, 2019, p. 15; also Datenethikkommission, Gutachten, 2019, p. 43, https://www.bmi.bund.de/SharedDocs/downloads/DE/publikationen/themen/it-digitalpolitik/gutachten-datenethikkommission.pdf?__blob=publicationFile&v=6.

[8] The AI Act applies with some exceptions from 2 August 2026 (Art. 113 AI Act).

[9] See also before Datenethikkommission, Gutachten, 2019, p. 163-64; Chapter 1 of the European Declaration on Digital Rights and Principles for the Digital Decade, 2023/C 23/01.

protect health, safety and fundamental rights (Art. 14 (2) AI Act). In order to effectively bind the AI system to such fundamental values of society, effective control[10] by members of such society (i.e. a human) seems promising.[11] Control goes hand in hand with and can be achieved through various concepts, e.g., the accountability of humans and/or governance mechanisms such as human oversight.

In the context of the AI Act, elements of human oversight can be found particularly with respect to AI-based decision-making. This makes sense from a policy perspective, as decisions[12] could have a particular impact on humans, unlike the use of AI as a mere tool for gathering information in less fundamental rights-sensitive contexts. These considerations are reflected in Art. 6 (2), (3) AI Act in conjunction with Annex III of the AI Act by considering AI systems as high-risk AI systems if they are used for decision-making[13] and pose a potential risk of "harm to the health, safety or fundamental rights of natural persons". Consequently, human oversight as a specific obligation under Art. 14 AI Act only applies to such high-risk AI systems.

When it comes to such decisions, it has to be distinguished between humans as decision-makers using AI systems (sub. a)) and the effects on humans on the other side as subjects of such decisions (sub. b)).

a) *Humans using AI systems for decision-making*

Humans, as organized in a company using AI systems for decision-making (i.e., providers and deployers), are required to comply with the human oversight obligations under

[10] See on the relationship of control and oversight Beck and Burri, From "Human Control" in International Law to "Human Oversight" in the New EU Act on Artificial Intelligence, in Mecacci et al. (eds.), Research Handbook on Meaningful Human Control of Artificial Intelligence Systems, 2024, p. 104, 104: "In other words, control is what the EU AI Act aims for globally, while oversight figures as just one component in the quest for human control of AI.".

[11] Cf. also Enqvist, 'Human oversight' in the EU artificial intelligence act: what, when and by whom?, Law, Innovation and Technology 15 (2023), 508, 511-15.

[12] Decision refers to the choice of different alternatives, Paal, Artikel 22 DS-GVO: Kreditscoring vor dem EuGH, ZfDR 2023, 114, 122; von Lewinski, in BeckOK DatenschutzR, 50th ed. 1 November 2024, Art. 22 DSGVO mn. 14 et seqq.

[13] Because Art. 6 (3) AI Act excludes many systems which are merely used as support tool for human decision-making, the criticism on Art. 14 AI Act addressing such support systems as well is obsolete, criticized before, e.g., by Bomhard/Merkle, Regulation of Artificial Intelligence, EuCML 2021, 257, 260.

the AI Act. Since machines are not legally responsible for themselves, this oversight obligation ensures the legal and moral[14] responsibility and contributes to the accountability of humans.[15] Unlike in other areas of law, the concept of accountability is less dependent on a specific human (or cooperation), and more on the idea of any human being responsible.[16] This mechanism also ensures that the policies set by a provider specifically determined for an AI system are consistently followed downstream.[17]

Humans on this side of decision-making have to ensure that human's autonomy is not undermined and that other humans are not adversely affected by a decision. In addition, challenges posed by the use of algorithms such as algorithmic bias[18] need to be addressed. In this light, the understanding of the automatic decision process and the (free)[19] possibility to disagree are core elements for effective human oversight.[20] Human oversight can be implemented through various approaches, such as human-in-the-loop (see below, sub. II.), and is required for the entire lifecycle of an AI system.[21]

b) Humans as subject of a decision

Already under the GDPR, and going back to 1978,[22] humans have the right not to be subject to an automated decision, which includes even more AI decision-making.[23] As

[14] Datenethikkommission, Gutachten, 2019, p. 14-15, 40; Dienes, Anforderungen an die menschliche Aufsicht über Künstliche Intelligenz, MMR 2024, 456, 461; cf. Nida-Rümelin, in Chibanguza, Kuß, and Steege (eds.), Künstliche Intelligenz, 1st ed. 2022, 1. Teil § 1 mn. 1 with further references.

[15] Constantino Torres, Exploring Article 14 of the EU AI proposal: accountability challenges of the human in the loop when supervising high-risk AI systems in public administration, 17 August 2022, p. 29, https://ssrn.com/abstract=4254940. However, not necessarily the accountability of controllers (as the focus is on humans), as argued under the GDPR by Lazcoz and de Hert, Computer Law & Security Review 50 (2023) 105833, 15.

[16] This is demonstrated by the reference to natural persons in Art. 26 (2) AI Act.

[17] On this triangular 'principal-agent' relationship Constantino Torres, Exploring Article 14 of the EU AI Proposal: Accountability Challenges of the Human in the Loop When Supervising High-Risk AI Systems in Public Administration, 17 August 2022, p. 20.

[18] E.g., Kuśmierczyk1, Algorithmic bias in the light of the GDPR and the proposed AI Act, in Morwaska and Olejnik (eds.), (In)equalities – Faces of modern Europe, 2023, p. 263; Recitals 27, 31, 56, 58, 67, 70 und 75 AI Act.

[19] I.e., by being aware of the automation bias, cf. Art. 14 (4) (b) AI Act.

[20] Beck and Burri, From "Human Control" in International Law to "Human Oversight" in the New EU Act on Artificial Intelligence, in Mecacci et al. (eds.), Research Handbook on Meaningful Human Control of Artificial Intelligence Systems, 2024, p. 104, 107-08.

[21] Recital 73 (1) AI Act.

[22] Including Art. 15 Data Protection Directive 95/46/EC and an even older French provision, as discussed by Martini, in Paal and Pauly (eds.), 3rd ed. 2021, Art. 22 DS-GVO mn. 14; Gola, in Bretthauer et al. (eds.), Verfassungen – ihre Rolle im Wandel der Zeit, 2019, p. 183, 196 et seq.

[23] See for the term "automated" Article 29 Working Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, WP 251rev.01, 6 February 2018, p. 8, https://ec.europa.eu/newsroom/article29/items/612053; on the policy Enquete-Kommission Kün-

the relevant Art. 22 GDPR has to be interpreted broadly in light of the ECJ's Schufa decision (see below, sub. 3.)[24] and thus covers many forms of integration of AI systems with substantial influence on decision-making processes, there was no need for such a provision under the AI Act. However, Art. 86 AI Act attaches legal consequences to AI-based decision-making on the side of the addressee of such a decision by providing for a right to explanation.[25]

3. Interplay with Art. 22 GDPR

Art. 22 GDPR does not only provide for a general prohibition[26] of automated decision-making in Art. 22 (1) GDPR with exceptions laid down in Art. 22 (2) GDPR,[27] but also provides for an individual right of natural persons to a human-made assessment of the decision, i.e., a human-based (new)[28] decision.

a) Scope of Art. 22 GDPR

The scope is limited to "a decision based solely on automated processing, including profiling, which produces legal effects concerning him or her or similarly significantly affects him or her", but also covers the transfer of "a probability value based on personal data relating to a person and concerning his or her ability to meet payment commitments in the future constitutes 'automated individual decision-making' within the meaning of that provision, where a third party, to which that probability value is transmitted, draws strongly on that probability value to establish, implement or terminate a contractual relationship with that person".[29] Similar to Art. 6 (2), (3) AI Act in con-

stliche Intelligenz – Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale, Unterrichtung, in BT-Drs. 19/23700, 83, https://www.bundestag.de/resource/blob/803528/01a4649b4f163856007be3371a2fe78e/Schlussbericht-der-Enquete-Kommission.pdf.

[24] CJEU, Case C-634/21, Schufa Holding, ECLI:EU:C:2023:957.

[25] In detail Radtke, Das Recht auf Erklärung unter der DSGVO und der KI-VO, in Dregelies, Henke and Kumkar (eds.), Artificial Intelligence: Rechtsfragen und Regulierung künstlicher Intelligenz im Europäischen Binnenmarkt, 9. Tagung GRUR Junge Wissenschaft, Nomos, 2025, p. 53.

[26] CJEU, Case C-634/21, Schufa Holding, ECLI:EU:C:2023:957, para 52.

[27] Those exceptions are related to the performance of a contract (lit. a), applicable Union or Member State law (lit. b), and the data subject's explicit consent (lit. c). On the distinction between Art. 22 (1), (2) GDPR Lazcoz and de Hert, Computer Law & Security Review 50 (2023) 105833, 9.

[28] Classifying this as measure ensuring human intervention *afterwards*, COM, White Paper on Artificial Intelligence – A European approach to excellence and trust, COM(2020) 65 final, 19 February 2020, p. 21, https://commission.europa.eu/document/download/d2ec4039-c5be-423a-81ef-b9e44e79825b_en?filename=commission-white-paper-artificial-intelligence-feb2020_en.pdf.

[29] CJEU, Case C-634/21, Schufa Holding, ECLI:EU:C:2023:957, para 73.

junction with Annex III AI Act, any substantial human influence on the specific decision precludes the application of Art. 22 GDPR.[30] However, in the case of substantial *human* influence, the typical risks of *automated* decisions do not materialize as they would in case of a solely automated decision.[31] This is because the typical risks are to be seen in particular in the objectification (cf. Art. 1 EU Charter) due to possible incomprehensible decisions of a machine that is not held accountable (see below, sub. b)). In practice, even this limited scope leaves many relevant decisions, and it would require a lot of effort by controllers if all of these decisions were to be challenged for human review.

If Art. 22 (1) GDPR applies and there is only a contract or data subject's consent but no specific Union or Member State law allowing for such an automated decision, the controller is under the obligation to implement safeguards including in any case "the right to obtain human intervention, on the part of the controller, to express his or her point of view and to contest the decision".

b) Subjective right in line with the relevant policy

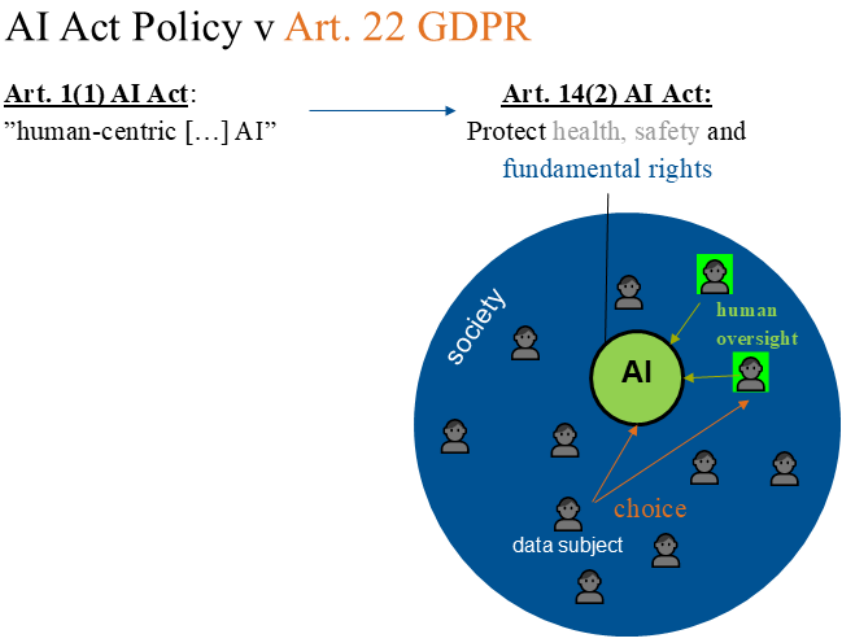
Of particular importance for data subjects who are affected by a decision within the scope of Art. 22 (1) GDPR are the individual rights with respect to such a decision. This includes a potential subjective right to have the decision assessed by a human being.

The nature as a subjective right to a human-made (new) decision is not clear from the wording of Art. 22 (3) GDPR, which only addresses the controller. However, taking into account the designation as a (general) "right" in Art. 22 (1) GDPR, the mandatory nature of further specified rights under Art. 22 (3) GDPR, the objective of the

[30] Article 29 Working Party, Guidelines on Automated individual decision-making and Profiling for the purposes of Regulation 2016/679, WP 251rev.01, 6 February 2018, p. 21.

[31] Contra Kuśmierczyk1, Algorithmic bias in the light of the GDPR and the proposed AI Act, in Morwaska and Olejnik (eds.), (In)equalities – Faces of modern Europe, 2023, p. 263, under IV. C.

provision and Recital 71 (1), (4) GDPR,[32] they clearly indicate a right of the data subject, regardless of whether the controller implements such (mandatory) safeguards or not.[33]



In providing such a right to data subjects, the policy under the GDPR can be summarized as rejecting AI-made decisions by giving the data subject the option to have such a decision later overridden by a human-made decision. The AI Act, however, accepts the concept of AI decision-making in general, but seeks to set the framework for

[32] „The data subject should have the right not to be subject to a decision [...]. In any case, such processing should be subject to suitable safeguards, which should include specific information to the data subject and the right to obtain human intervention, to express his or her point of view, to obtain an explanation of the decision reached after such assessment and to challenge the decision."

[33] See also CJEU, Case C-634/21, Schufa Holding, ECLI:EU:C:2023:957, para 52: "the infringement of which *does not need* to be invoked individually by such a person" (emphasis added), which does not entails that such infringement *cannot* be invoked by a data subject; with a similar result as here Martini, in Martini and Wendehorst (eds.), 1st ed. 2024, Art. 14 KI-VO mn. 15.

the best possible AI-made decision through design decisions that take into account human oversight.[34]

These different approaches make sense from a policy perspective: Art. 22 GDPR, on the one hand, aims primarily to protect against the objectification of humans by a machine due to limited controllability and traceability, with implications for human dignity (Art. 1 EU Charter), and secondarily to protect against machine decisions based on inaccurate data (e.g., under Art. 16 GDPR).[35] Those data subjects who are particularly affected by an AI-made decision can use the right under Art. 22 (1), (3) GDPR as opt-out of the automated decision. The AI Act, on the other hand, does not place a particular emphasis on the risk of objectification, but rather focuses on optimizing the AI decision-making process and enabling the exercise of rights under other legal acts.[36]

II. Classification of measures

The AI HLEG proposes a classification of mechanisms to human governance[37] such as human-in-the-loop (HITL), human-on-the-loop (HOTL), and human-in-command (HIC).[38] This classification is not strictly followed in the literature,[39] but can still provide some guidance for system design decisions on implementing human oversight and thus shed light on the human oversight requirement under the AI Act. These classifications can be further divided into measures built into the AI system and measures to be implemented by the deployer.[40]

---

[34] Radtke, Das Verhältnis von KI-VO und Art. 22 DSGVO unter besonderer Berücksichtigung der Schutzzwecke, RDi 2024, 353, 357.

[35] In detail Radtke, Das Verhältnis von KI-VO und Art. 22 DSGVO unter besonderer Berücksichtigung der Schutzzwecke, RDi 2024, 353, 355-57.

[36] E.g., Recitals 9 (2), 171 (2) AI Act; in detail Radtke, Das Verhältnis von KI-VO und Art. 22 DSGVO unter besonderer Berücksichtigung der Schutzzwecke, RDi 2024, 353, 357-359. Such other legal acts may include product liability and tort law. See below for the liability under the Product Liability Directive and the AI Liability Directive (Draft), sub. III. 3.

[37] E.g., Lazcoz and de Hert, Computer Law & Security Review 50 (2023) 105833, 8.

[38] HLEG-AI, Ethics Guidelines for Trustworthy AI, 2019, p. 16; see also Enquete-Kommission Künstliche Intelligenz – Gesellschaftliche Verantwortung und wirtschaftliche, soziale und ökologische Potenziale, Unterrichtung, BT-Drs. 19/23700, 83 fn. 250; additional classification by Van Dijck, Predicting recidivism risk meets ai act. European Journal on Criminal Policy and Research 28 (2022), 407, 419-20: optional, benchmark, and feedback approach.

[39] E.g., Constantino Torres, Exploring Article 14 of the EU AI proposal: accountability challenges of the human in the loop when supervising high-risk AI systems in public administration, 17 August 2022, p. 29-30.

[40] Art. 14 (3) AI Act.

1. Human-in-the-loop (HITL)

The AI HLEG defines HITL as "the capability for human intervention in every decision cycle of the system".[41] Interestingly, the AI HLEG acknowledges that such an approach is "in many cases neither possible nor desirable".[42] This advice appears to have been followed by the European legislator as arguably no such approach has been implemented under the AI Act.

However, Art. 22 GDPR generally follows such an approach by giving the data subject the right to a human-made (new) decision and thus to "human intervention" (see above, sub. I. 3.).[43] In cases where human intervention is implemented and not only upon request carried out for every decision, the AI system would often not be considered a high-risk AI system within the meaning of Art. 6 (2), (3) AI Act and thus the human oversight obligation under Art. 14 AI Act would not apply. In other words: A consistent implementation of HITL can be a way out of the scope of the AI Act,[44] but not a concept within the AI Act (with the exception of Art. 6 (3) AI Act).

2. Human-on-the-loop (HOTL)

HOTL approaches focus on the design and monitoring levels. Accordingly, the AI HLEG describes HOTL as the "capability for human intervention during the design cycle of the system and monitoring the system's operation".[45] Elements of the AI Act follow the HOTL approach by requiring technical (design) measures[46] and are thus mostly preventive in nature.[47]

The technical documentation of the AI system (Art. 11 AI Act), the information for the deployer (Art. 13 AI Act) and the built-in mechanisms to create records of the use of the AI system (Art. 12, 19 AI Act) are design measures to be implemented primarily by the provider. Taken together, such information about the AI system and the particular decision process provides individuals with the system-specific know-how and the ability to monitor the system's operation in the sense of HOTL.

---

[41] HLEG-AI, Ethics Guidelines for Trustworthy AI, 2019, p. 16.
[42] HLEG-AI, Ethics Guidelines for Trustworthy AI, 2019, p. 16.
[43] Similar Lazcoz and de Hert, Computer Law & Security Review 50 (2023) 105833, 2.
[44] However, Ch. IV for AI systems and Ch. V for AI models could still apply.
[45] HLEG-AI, Ethics Guidelines for Trustworthy AI, 2019, p. 16.
[46] Considering the approach of the AI Act as management-based oversight of AI Coglianese and Crum, Taking Training Seriously: Human Guidance and Management-Based Regulation of Artificial Intelligence, Public Law and Legal Theory Research Paper Series Research Paper No. 24-08 p. 4, https://arxiv.org/pdf/2402.08466.
[47] Enqvist, 'Human oversight' in the EU artificial intelligence act: what, when and by whom?, Law, Innovation and Technology 15 (2023), 508, 518.

3. Human-in-command (HIC)

HIC lies in between HITL and HOTL in that it is based on the decision to use the AI system for a particular decision or in a particular situation, taking into account the impact of the AI system on a broader level. The AI HLEG defines HIC as "capability to oversee the overall activity of the AI system (including its broader economic, societal, legal and ethical impact) and the ability to decide when and how to use the system in any particular situation".[48]

The obligation to design the AI system on the basis of a risk management process (Art. 9 AI Act) and to take into account specific risks with regard to data governance (Art. 10 (2) (f) AI Act) as well as a fundamental rights impact assessment in specific cases (Art. 27 AI Act) ensures that the broader overall societal impact of the AI system is taken into account when developing the AI system. Furthermore, the implementation of the 'stop'-button under Art. 14 (4) (e) AI Act and at the operation level, the decision not to use an AI system in a particular situation (Art. 14 (4) (d) AI Act) or to use such a 'stop'-button in a particular situation are HIC governance mechanisms. It is the responsibility of the deployer and the human overseers to make use of these mechanisms on a case-by-case basis taking into account the information provided by the provider of the AI system (e.g., pursuant to Art. 13 AI Act).

4. No-human-in-the-loop?

In legal scholarship, it has been argued that in some situations a "no human in the loop" approach may be more effective and thus preferable.[49] This is due to further to be examined challenges of the concept of human oversight in general (see below, sub. IV.) and the potential effectiveness of AI decision-making, e.g., in the health sector.[50] While not only Art. 14 (4) AI Act, but also fundamental rights as laid down in the EU Charter, require that human oversight does not significantly impair the effectiveness of AI systems,[51] disadvantages resulting from the integration of human influence (such as negative effects of discretion)[52] must be accepted in light of the strict requirement under

[48] HLEG-AI, Ethics Guidelines for Trustworthy AI, 2019, p. 16; BT-Drs. 19/23700, 83 fn. 250.

[49] Gassner, Menschliche Aufsicht über intelligente Medizinprodukte, MPR 2023, 5, 10.

[50] Chockley and Emanuel, The End of Radiology? Three Threats to the Future Practice of Radiology, Journal of the American College of Radiology 13 (2016), 1415, 1417-19; Nicholson Price II, Artificial Intelligence in Health Care Applications and Legal Issues, Scitech Law. 14 (2017), 10; Selbst, Negligence and ai's human users, Boston University Law Review 2020 (100), 1315, 1334-35; Gassner, Menschliche Aufsicht über intelligente Medizinprodukte, MPR 2023, 5, 10-11.

[51] Gassner, Menschliche Aufsicht über intelligente Medizinprodukte, MPR 2023, 5, 11.

[52] See for the positive effects Green, The flaws of policies requiring human oversight of government algorithms, Computer Law & Security Review 45 (2022) 105681, 5-6 under 3.2 with further references.

Art. 14 AI Act. This is particularly true with respect to human dignity (Art. 1 EU Charter)[53] for correct, but non-transparent AI-made decisions (see above, sub. I. 3. b)).[54]

5. Results

One could say that the GDPR aims at human-in-the-loop (i.e. "human intervention in every decision cycle of the system"), while the AI Act requires less control over the result, but more over the process, which could be classified as human-on-the-loop and human-in-command concepts. However, a no-human-in-the-loop concept is not warranted under the AI Act.

III. Human oversight under the AI Act

Human oversight obligations apply both to the provider under Art. 16 (a) in conjunction with Art. 14 AI Act and to the deployer under Art. 26 (2) AI Act, each of which obligations is subject to a fine pursuant to Art. 99 AI Act.

1. Human oversight by design

The provider is obliged to carry out a risk-based assessment and to implement necessary technical and organizational measures to enable effective human oversight, e.g., by building measures into the AI system or identifying measures to be implemented[55] by the deployer (Art. 14 (3) AI Act, see above, sub. II. 2. and 3.). The requirements depend on the effectiveness of these measures for human oversight during the operation of the AI system. Despite the fact that the requirements were diluted during the legislative

[53] Cf. Lazcoz and de Hert, Computer Law & Security Review 50 (2023) 105833, 13; Mendoza and Bygrave, The Right not to be Subject to Automated Decisions based on Profiling, University of Oslo Faculty of Law Legal Studies Research Paper Series No. 2017-20, p. 7, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2964855.

[54] In detail Radtke, Das Verhältnis von KI-VO und Art. 22 DSGVO unter besonderer Berücksichtigung der Schutzzwecke, RDi 2024, 353, 355-56.

[55] Highlighting that this provision nevertheless addresses the providers and not the deployers Enqvist, 'Human oversight' in the EU artificial intelligence act: what, when and by whom?, Law, Innovation and Technology 15 (2023), 508, 518.

process,[56] this threshold can still be considered high as such measures by the provider[57] should enable humans overseeing the AI systems to understand capacities and limitations, remain aware of the automation bias, correctly interpret (but not necessarily: understand) the output, and decide not to use or to stop an AI system (Art. 14 (4) AI Act).[58] While these goals may contribute to a human-centric AI system, if implemented, the devil is in the implementation of measures to achieve the outlined goals (see below, sub. IV.).

Other measures at the design level of the AI systems contribute to an environment for human oversight. These include the technical documentation (Art. 11, Annex IV (2) (e) AI Act), instructions for the deployers (Art. 13 (3) (d), Recital 72 (5) AI Act)[59] and the fundamental rights impact assessment (Art. 27 (1) (e) AI Act), which serve as a basis for effective human oversight.

2. Assignment of human oversight to natural persons

The deployer is obliged to actually assign human oversight to natural persons and to select such persons (Art. 26 (2) AI Act).[60] This can be considered as a technical and organizational measure complementing those of the provider under Art. 14, 16 AI Act.[61]

It is not specified whether such natural persons must be employed by the deployer or may be employed by the provider, which could make sense if the provider retains control over oversight mechanisms such as the system logs (cf. Art. 16 (e) AI Act). From the perspective of the AI Act, it is less important, to whom the natural persons are assigned, but rather – in line with human oversight as a tool to human-centric AI – whether they are natural persons and have the necessary qualifications.

Accordingly, Art. 26 (2) AI Act contains a broad catalog of requirements for the selection of the natural persons for human oversight: They should have the „necessary compentence, training, [...] authority [...] and support". The systematic of the AI Act helps to shed some light on these requirements. AI Literacy under Art. 4 AI Act can be

---

[56] Enqvist, 'Human oversight' in the EU artificial intelligence act: what, when and by whom?, Law, Innovation and Technology 15 (2023), 508, 519-20; Dienes, Anforderungen an die menschliche Aufsicht über Künstliche Intelligenz, MMR 2024, 456, 459-60; Gassner, Menschliche Aufsicht über intelligente Medizinprodukte, MPR 2023, 5, 5-6.

[57] Emphasizing that only the provider is addressed Enqvist, 'Human oversight' in the EU artificial intelligence act: what, when and by whom?, Law, Innovation and Technology 15 (2023), 508, 522.

[58] Agreeing with the essence of these requirements Constantino Torres, Exploring Article 14 of the EU AI proposal: accountability challenges of the human in the loop when supervising high-risk AI systems in public administration, 17 August 2022, p. 31-34.

[59] Lazcoz and de Hert, Computer Law & Security Review 50 (2023) 105833, 7.

[60] Recitals 73, 91 (3) AI Act.

[61] Eisenberger, in Martini/Wendehorst (eds.), 1st ed. 2024, Art. 26 KI-VO mn. 25.

seen as the first level, and human oversight competence under Art. 26 (2) AI Act as the second level.

The AI literacy requirement in Art. 4, Art. 3 (56) AI Act applies to all providers and deployers of any AI system, i.e. not only to high-risk AI systems. It is not specific to an AI system,[62] but requires "a sufficient level of [general] AI literacy", taking into account the context of the "AI systems" of the provider or deployer.[63] This general understanding of AI systems lays the foundation for human oversight competence with higher requirements.

The human oversight competence goes beyond AI literacy and takes into account the AI system in question, as can be concluded from the comparison with Art. 4 AI Act. The provider's obligation to implement measures to achieve the goals under Art. 14 (4) AI Act necessarily implies that the natural persons do not impede the achievement of these goals (e.g., by not applying the tools provided by the provider correctly). The degree of competence must be sufficiently high to ensure effective oversight within the meaning of Art. 14 (1), Art. 26 (2) AI Act. Thus, Art. 14 (4) AI Act can be used to concretize the competence requirements for natural persons and requires: an understanding of "the relevant capacities and limitations of *the* high-risk AI system" (emphasis added), awareness of the "automation bias", the correct interpretation of the output (where necessary to carry out oversight), and the decision not to use or stop the use of an AI system in a specific case. Competence should be taught and maintained through "training".

Along with the competence, the natural persons must have sufficient authority and support. This addresses the actual power to apply the skills derived from their competence.

3. Liability of the provider and deployer

Both the provider and the deployer of an AI system are liable for fines under Art. 99 AI Act if they violate their specific human oversight obligations as discussed above. These specific obligations arguably do not include the flawless performance of human oversight by natural persons. Under the AI Act, neither the deployer nor the natural person

[62] Nevertheless, it requires providers to take into account the context of the provider's AI systems, see also Fleck, AI literacy als Rechtsbegriff, KIR 2024, 99, 100-01. Contra Dienes, Anforderungen an die menschliche Aufsicht über Künstliche Intelligenz, MMR 2024, 456, 458; seeing Art. 4 as basis for the competence for human overseers Wendehorst, in Martini/Wendehorst (eds.), 1st ed. 2024, Art. 4 KI-VO mn. 5.

[63] Recital 20 AI Act; cf. Möller-Klapperich, Die neue KI-Verordnung der EU, NJ 2024, 337, 339-40.

is responsible for failures in human oversight,[64] provided that there were no failures in the selection of the natural person and the design of human oversight.[65] It could be argued that *natural* persons should be selected as representatives of core values including the discretion of humankind when overriding AI-based decisions.

However, national contract or tort law may provide for the liability of providers and deployers for errors of AI systems, including errors that could have been *remedied* by effective human oversight. In this context, and in particular for errors despite effective human oversight, such laws can be seen as a manifestation of the limitations of human oversight.[66] The recently amended Product Liability Directive (EU) 2024/2853 (PLD) requires changes in national law in this respect. According to its Art. 4 (1) it covers defective software including AI systems as can be deduced from Recital 3 (1) PLD. The focus of the liability is on the defect of a product and not so much on the procedure in relation to this product (e.g., human oversight). However, this liability regime is still relevant in light of human oversight: First, effective human oversight can be considered a "relevant product safety requirement" and thus relevant under Art. 7 (2) (f) PLD. Second, effective human oversight can prevent damage in individual cases and serve as a tool to avoid liability under the damage prong.

Beyond this framework, it could be even hard to establish liability for damages not for the performance of human oversight, but for the failure to comply with the obligations under Art. 14, 26 AI Act. For example, under German law, such failure could generally trigger liability for violation of Art. 14, 26 AI Act as a protective standard under § 823 (2) German Civil Code [BGB].[67] The necessary causal link between the violation of the standard and the damage is not only based on an omission, but would also require an assumption as to how the natural persons in charge of human oversight would have reacted. While such concepts are not uncommon in civil law,[68] such an assumption conflicts with the particular idea of human autonomy and the discretion of natural persons in the process. In other words: One cannot simply assume the "how" of the actual human oversight in a specific case, because human oversight serves the purpose of reflecting the individuality of human beings and the discretion of natural persons within the framework of the AI Act (e.g., regarding the competence and power of the natural persons under Art. 26 (2) AI Act).

---

[64] Dienes, Anforderungen an die menschliche Aufsicht über Künstliche Intelligenz, MMR 2024, 456, 458.

[65] Insofar this concepts reminds of the liability for the failure in selection of a vicarious agent under § 831 German Civil Code [BGB].

[66] Cf. Messner-Kreuzbauer and Pehm, Taming AI Through Presumptions: a Softer Approach to Tort Law Harmonisation?, ZEuP 2024, 161, 165-66 with further references.

[67] Overview on the discussions Theis, Auswirkungen der KI-Verordnung und der KI-Haftungs-Richtlinie auf die Haftung beim KI-Einsatz in der Finanzbranche, BKR 2024, 414, 415 with further references; Wendehorst, in Martini/Wendehorst (eds.), 1st ed. 2024, Art. 1 KI-VO mn. 85.

[68] E.g., for the Presumption of correct behaviour in the event of information [Vermutung aufklärungsrichtigen Verhaltens] Bundesgerichtshof, Case XI ZR 586/07, NJW 2009, 2298.

It remains to be seen what approach the AI Liability Directive, which has yet to be agreed and adopted, will take in this respect. In general, the draft of the AI Liability Directive covers a wide(r) range of damages, provides for the disclosure of evidence and a limited presumption[69] and would therefore remain relevant despite the new Product Liability Directive.[70] Recently, a presumption of causation between ex-post oversight and harmful outputs of AI systems has been proposed for the AI Liability Directive.[71] If the AI Liability Directive were to follow such an approach, the AI Liability Directive would provide strong incentives for effective human oversight, but could also raise questions about an unwarranted restriction of the discretion of natural persons overseeing AI systems.

IV. Challenges for effective human oversight

While the general idea of human oversight, may seem – and actually could be – promising, in practice, there are several challenges.

1. Incompatible concepts

Scholars have already argued that *consistent* AI-based decision-making and *flexible* human decision-making and oversight are incompatible concepts.[72] Adopting this line of reasoning, the AI Act and implemented human oversight may create a false sense of security among policymakers, AI system providers and deployers, potentially leading to the unwarranted use of AI systems for riskier purposes.[73]

[69] Art. 3 (1), 4 (1) of the AI Liability Directive Draft (COM/2022/496 final); European Parliamentary Research Service, Complementary Impact Assessment, September 2024, p. 14, https://www.europarl.europa.eu/RegData/etudes/STUD/2024/762861/EPRS_STU(2024)762861_EN.pdf.

[70] Messner-Kreuzbauer and Pehm, Taming AI Through Presumptions: a Softer Approach to Tort Law Harmonisation?, ZEuP 2024, 161, 167.

[71] European Parliamentary Research Service, Complementary Impact Assessment, September 2024, p. II: "The direct causation between a lack of ex-post oversight and harmful outputs is not always clear. It is suggested to establish a direct presumption of causality between AI outputs and damages for non-compliance with monitoring obligations".

[72] Cf. Green, The flaws of policies requiring human oversight of government algorithms, Computer Law & Security Review 45 (2022) 105681, 7-8, 12.

[73] Green, The flaws of policies requiring human oversight of government algorithms, Computer Law & Security Review 45 (2022) 105681, 9, who thus prefers institutional oversight on whether to use an algorithm in a particular case.

Not only is it challenging to *effectively* oversee a decision that is based on an completely different and difficult to understand mechanism,[74] but the combination of such different concepts raises questions about the clear allocation of responsibility in practice[75] and can be costly and ineffective.[76] This takes into account shortcomings of humans and their discretion, as well as the flaws of algorithms when they are not always accurate[77] and cannot account for human emotions.[78]

This challenge is illustrated by parallel discussions on the right to explanation and interpretability of automated decision-making and explainable AI.[79] However, it is still unclear whether and to what extent the legislator's desire for comprehensible explanations of AI decision-making processes, as expressed in Art. 86 AI Act, can be fulfilled in practice.

2. Broad and far-reaching requirements under Art. 14, 26 AI Act

Assuming that AI decision-making and human oversight are generally compatible, there are challenges in the details of the human oversight requirements. As has been shown above (sub. III. 1., 2.), Art. 14, 26 AI Act provide for far-reaching, high standards,[80] including the provision of the necessary means to overcome potential flaws of the human-machine interaction (e.g., the automation bias and interpretability issues).

---

[74] Green, The flaws of policies requiring human oversight of government algorithms, Computer Law & Security Review 45 (2022) 105681, 7 with further references; on preconditions to effective human oversight in general Article 36, Key elements of meaningful human control, Background paper to comments prepared by Richard Moyes, Managing Partner, Article 36, for the Convention on Certain Conventional Weapons (CCW) Meeting of Experts on Lethal Autonomous Weapons Systems (LAWS), April 2016, p. 3-4, https://www.article36.org/wp-content/uploads/2016/04/MHC-2016-FINAL.pdf.

[75] Beck and Burri, From "Human Control" in International Law to "Human Oversight" in the New EU Act on Artificial Intelligence, in Mecacci et al. (eds.), Research Handbook on Meaningful Human Control of Artificial Intelligence Systems, 2024, p. 104, 108; cf. Green, The flaws of policies requiring human oversight of government algorithms, Computer Law & Security Review 45 (2022) 105681, 10-11.

[76] Cf. Constantino Torres, Exploring Article 14 of the EU AI proposal: accountability challenges of the human in the loop when supervising high-risk AI systems in public administration, 17 August 2022, p. 33 with further references.

[77] Green, The flaws of policies requiring human oversight of government algorithms, Computer Law & Security Review 45 (2022) 105681, 3-4; Hildebrandt, Algorithmic regulation and the rule of law, Philosophical Transactions R. Soc. A 376 (2018) 20170355, 3, http://dx.doi.org/10.1098/rsta.2017.0355.

[78] Addressing this issue Yoo and Jeong, EP-bot: Empathetic chatbot using auto-growing knowledge graph, Computers, Materials, & Continua *67* (2021), 2807, 2808, https://doi.org/10.32604/cmc.2021.015634.

[79] Radtke, Das Recht auf Erklärung unter der DSGVO und der KI-VO, in Dregelies, Henke and Kumkar (eds.), Artificial Intelligence: Rechtsfragen und Regulierung künstlicher Intelligenz im Europäischen Binnenmarkt, 9. Tagung GRUR Junge Wissenschaft, Nomos, 2025, p. 53; Dienes, Anforderungen an die menschliche Aufsicht über Künstliche Intelligenz, MMR 2024, 456, 461; Holzinger et al. (eds.), xxAI - Beyond Explainable AI, 2020.

[80] Van Dijck, Predicting recidivism risk meets ai act, European Journal on Criminal Policy and Research 28 (2022), 407, 419; Enqvist, 'Human oversight' in the EU artificial intelligence act: what, when and by whom?, Law, Innovation and Technology 15 (2023), 508, 509, 532.

These approaches, while plausible in theory, may have limitations in practice. For example, whether the automation bias can be overcome depends in particular on (effective) methods to develop and to implement.[81] Overcoming the automation bias goes along with understanding AI-based decisions. This requires not only tools such as interpretable and explainable AI,[82] but also resources in terms of training and financial resources.[83] In addition, the ability of humans to effectively oversee AI-made decisions may be generally limited by the human body[84] or by the use of systems in fields where software even sets the baseline for the standards to be applied (e.g., for the concept of normal traffic in the network security context).[85]

However, as the discussion on explainable AI shows, it cannot be ruled out that the solution lies in the (further) development of the technology itself. In addition, other conceptual improvements, such as providing feedback[86] to the natural persons overseeing the AI systems, may be promising to tackle some of the challenges discussed above.

3. Interplay with other legal acts, in particular Art. 22 GDPR

On the face of it, the interaction of the AI Act with other legislation increases the complexity of the legal assessment and could hamper effective human oversight through parallel concepts in other legislation and uncertainty about the relationship of application. In particular, according to Art. 2 (7) AI Act,[87] the GDPR remains unaffected without any elaboration.

However, a closer look suggests that the AI Act and Art. 22 GDPR actually work well together, taking into account the different policy considerations (see above, sub. I. 3. b)). The AI Act aims at the best possible AI-based decision process,[88] taking into account the focus on human-centric AI. Rights such as the right to explanation under

[81] Constantino Torres, Exploring Article 14 of the EU AI proposal: accountability challenges of the human in the loop when supervising high-risk AI systems in public administration, 17 August 2022, p. 42-44.

[82] Explainable AI may allow for the explanation, while Art. 14 (4) AI Act only requires interpretability, as argued by Dienes, Anforderungen an die menschliche Aufsicht über Künstliche Intelligenz, MMR 2024, 456, 461.

[83] Constantino Torres, Exploring Article 14 of the EU AI proposal: accountability challenges of the human in the loop when supervising high-risk AI systems in public administration, 17 August 2022, p. 39-42.

[84] Baxter et al., The ironies of automation … still going strong at 30?, 2012, p. 65-67, https://dl.acm.org/doi/pdf/10.1145/2448136.2448149.

[85] Selbst, Negligence and ai's human users, Boston University Law Review 2020 (100), 1315, 1337.

[86] Biermann, Horton, and Walter, Algorithmic Advice as a Credence Good, ZEW Discussion Paper No. 22-071, 2022, p. 10 et seqq., https://ftp.zew.de/pub/zew-docs/dp/dp22071.pdf.

[87] See also Recitals 9 (2), 10 (1), Art. 59 (3) AI Act.

[88] On the phases of development and use of an AI system under the AI Act and the GDPR Lazcoz and de Hert, Computer Law & Security Review 50 (2023) 105833, 6.

Art. 86 AI Act facilitate the enforcement of rights provided for in other legal acts,[89] e.g., under national tort law. In cases where this combination of HOTL and HIC approach does not work out and data subjects feel objectified due to non-transparent decisions, they have the right[90] to a human-made assessment of a *specific* decision, i.e., at the stage of AI use rather than the stage of development. Arguments in favor of a stronger actual human influence under Art. 14 AI Act[91] are thus at least to be partially rejected by reference to Art. 22 GDPR.

Although there are challenges in the interaction between the two legal acts, the different approaches could prove to be promising when taken together.

V. Summary and outlook

From a mere policy perspective, the concept of human oversight under the AI Act fits perfectly with the objective of human-centric AI (Art. 1 (1) AI Act). While Art. 22 GDPR follows a human-in-the-loop approach to address concerns about objectifying humans with respect to human dignity, the AI Act implements human-on-the-loop and human-in-command structures to ensure value-bound AI and to enable rights under other legal acts. Arguably, both legal acts could work well together enabling unsatisfied data subjects to exercise their rights under Art. 22 (1), (3) GDPR.

However, the devil is in the details: The implementation of the human oversight obligations outlined in Art. 14, 26 AI Act poses significant practical challenges, both in terms of conceptual frameworks and the interpretability and explainability of decisions made by AI systems. The effectiveness of human oversight in overcoming these challenges remains to be explored. Potential solutions may be found in further technological development or in the effective interplay between the different legal concepts set out in Art. 22 GDPR and the AI Act. The liability regime established under the PLD and potentially complemented by an AI Liability Directive, may incentivize providers and deployers of AI systems to ensure effective human oversight in terms of decision outcomes.

---

[89] Recitals 9 (2), 171 (2) AI Act.

[90] Highlighting the different natures of Art. 14 AI Act as an obligation of the providers and Art. 22 GDPR providing individuals with a subjective right Martini, in Martini/Wendehorst (eds.), 1st ed. 2024, Art. 14 KI-VO mn. 15.

[91] E.g., Beck and Burri, From "Human Control" in International Law to "Human Oversight" in the New EU Act on Artificial Intelligence, in Mecacci et al. (eds.), Research Handbook on Meaningful Human Control of Artificial Intelligence Systems, 2024, p. 104, 114; see also on the risk of a low threshold of meaningful human interaction Wagner, Liable, but Not in Control? Ensuring Meaningful Human Agency in Automated Decision-Making Systems, Policy & Internet 11 (2019), 104, 114-16.

This to say, with regard to human oversight as a core principle under the AI Act, interesting times lie ahead before the AI Act becomes generally applicable on 2 August 2026.

The Regulatory Approach of the European Union's Artificial Intelligence Act

David Restrepo Amariles[1], Aurore Troussel[2]

The European Union's Artificial Intelligence (AI) Act establishes a comprehensive, horizontal regulatory framework for artificial intelligence, marking a transformative moment in global AI governance. This article critically examines the Act's regulatory approach and assesses its prospective effectiveness. First, we conceptualize the AI Act as a hybrid model of meta-regulation that integrates human rights and product safety regulation through risk-based, principles-based, and rights-based approaches, moving beyond conventional command-and-control mechanisms. To elucidate its distinctive model of meta-regulation, we situate the AI Act within the broader EU regulatory landscape, drawing specific comparisons with the General Data Protection Regulation (GDPR) and the Digital Services Act (DSA). We argue that the AI Act aims to balance technological innovation and fundamental rights protection by relying heavily on self-assessment and on co-regulation tools, like sandboxes. This approach appears to offer flexibility to providers and deployers, even if it increases their compliance burden and may raise concerns about enforcement consistency and the effectiveness of compliance mechanisms. The article concludes that the AI Act puts forward an ambitious and novel regulatory paradigm, but its long-term effectiveness will depend on the adaptability of its institutional framework and regulatory mandates in response to evolving market dynamics and technological advancements.

I. Introduction

The AI Act constitutes a groundbreaking regulation, introducing harmonized rules for artificial intelligence (AI) across the European Union (EU). As the first legally binding, horizontal, and comprehensive AI regulation globally, it has set a precedent for other

[1] HEC Paris.
[2] University of Montreal & HEC Paris.

jurisdictions—both positively and negatively[3]—as they seek to establish their own approaches to AI governance[4]. In the absence of a global consensus or a binding international treaty governing artificial intelligence[5], the AI Act has emerged as a pivotal benchmark, influencing policy discourse among national governments and international organizations. Its prominence underscores the necessity of a rigorous examination of its underlying regulatory approach and an assessment of its potential effectiveness.

The regulation of AI presents complex and multifaceted challenges[6]. Excessive regulatory intervention risks stifling innovation and undermining a country's competitiveness in the global AI landscape. Conversely, insufficient regulation may expose individuals, businesses, and society at large to significant risks, including threats to fundamental rights, market integrity, and systemic stability. As AI systems become increasingly integrated into commercial and public applications, a growing consensus has emerged regarding the necessity of a regulatory framework[7]. However, the precise nature of such regulatory framework remains contested, with approaches varying across jurisdictions—ranging from ethical guidelines, codes of conduct, technical standards, and risk-based governance models to more prescriptive, state-driven regulatory regimes[8].

However, it is important to recognize that AI deployment, particularly in highly regulated sectors such as healthcare, finance, and insurance, has not taken place in a legal vacuum while awaiting the development of specific AI regulations. General product safety and data protection laws already impose regulatory constraints on AI across many jurisdictions, including the EU, while sector-specific regulations introduce addi-

[3] Veale and Zuiderveen Borgesius, Demystifying the Draft EU Artificial Intelligence Act — Analysing the good, the bad, and the unclear elements of the proposed approach, Computer Law Review International, vol. 22, no. 4, 2021, 97-112.

[4] See: Reed, How should we regulate artificial intelligence? Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences, 2018 ; Clarke, Regulatory alternatives for AI, Computer Law & Security Review, 2019, 398-409 ; Almeida, dos Santos & Farias, Artificial Intelligence Regulation: a framework for governance, Ethics Inf Technol 23, 2021, 505–525.

[5] There is, however, the notable development of the Council of Europe Framework Convention on Artificial Intelligence and Human Rights, Democracy, and the Rule of Law, which was opened for signature on 5 September 2024.

[6] Feldman, Lemley, Masur, Rai, Open letter on ethical norms in intellectual property scholarship, Harvard Journal of Law & Technology, 2016.

[7] For a discussion on global consensus on AI regulation see: Restrepo Amariles, Regulating Artificial Intelligence – Is Global Consensus Possible?, Forbes, 9 September, 2022, https://www.forbes.com/sites/hecparis/2022/09/09/regulating-artificial-intelligence--is-global-consensus-possible/

[8] Ebers, Truly risk-based regulation of artificial intelligence how to implement the EU's AI Act, European Journal of Risk Regulation, 2024, 1-20 ; Gornet & Maxwell, The European approach to regulating AI through technical standards, Internet Policy Review 13.3, 2024, 1-27 ; Novelli, et al., AI Risk Assessment: A Scenario-Based, proportional methodology for the AI act, Digital Society 3.1, 2024, 13.

tional governance requirements for specific AI systems. In financial services, for instance, firms offering investment-related services must comply with the Markets in Financial Instruments Directive (MiFID) and the Market Abuse Regulation (MAR)[9], both of which establish stringent compliance obligations applicable to algorithmic trading services. These existing legal frameworks underscore the extent to which the AI Act is embedded within a broader constellation of regulatory instruments, challenging the notion that AI regulation in the EU is entirely novel or that it was needed to fill a regulatory vacuum.

This article examines the key structural elements that define the regulatory approach of the AI Act as a comprehensive and horizontal framework for AI regulation. It then conceptualizes the AI Act's regulatory approach as a form of meta-regulation and situates its approach within the broader landscape of EU digital regulations, including the General Data Protection Regulation (GDPR) and the Digital Services Act (DSA). Finally, it explores the implications of this regulatory approach for the Act's potential effectiveness and compliance dynamics.

II. Key Features of the AI Act

The AI Act is an integral part of the EU Digital Strategy and seeks to regulate AI systems based on their use and deployment within the EU market. The Act entered into force in August 2024, with most provisions applying after a two-year implementation period. However, prohibitions of certain AI systems will apply from February 2025, and requirements for General Purpose AI (GPAI) models will take effect in August 2025. The AI Act establishes horizontal rules for the regulation of AI systems, categorizing them based on their potential risks and uses. Providers and deployers of AI systems are subject to different obligations depending on the classification of the system. Notably, certain AI practices are outright prohibited due to their inherent risks, while high-risk AI systems must undergo a conformity assessment before being placed on the market. The AI Act consists of thirteen chapters, and follows the standard structure of European regulations. Four chapters specifically address different categories of AI systems:

[9] Restrepo Amariles & Lewkowicz, Unpacking smart law: how mathematics and algorithms are reshaping the legal code in the financial sector, Lex Electronica 25, 2020 ; Fortes, Borges, Marcello Baquero, & Restrepo Amariles, Artificial intelligence risks and algorithmic regulation, European Journal of Risk Regulation 13.3, 2022, 357-372 ; Tilen & Van Waeyenberge, European legal framework for algorithmic and high frequency trading (Mifid 2 and MAR): A global approach to managing the risks of the modern trading paradigm, European Journal of Risk Regulation 9.1, 2018, 146-153.

prohibited practices (Chapter II), high-risk AI systems (Chapter III), transparency requirements for certain AI systems (Chapter IV), and general-purpose models (Chapter V). These chapters establish rules applicable to all targeted AI systems, while AI systems not falling within these categories are excluded from regulation under the AI Act.

The AI Act has three interesting key features in relation to its regulatory approaches, namely (1) an extraterritorial scope, (2) an asymmetric regulation, and (3) a product-safety inspired regime complemented by sandboxes.

1. Extraterritorial scope

The AI Act applies not only to AI systems developed within the EU but also to systems placed on the EU market or put into service within EU member states, irrespective of whether the providers are located in the EU or third countries (Article 2.1(a)). This extraterritorial scope mirrors the approach of the EU's General Data Protection Regulation (GDPR). Furthermore, the Act applies to AI systems developed in third countries if their output is used within the EU (Article 2.1(c)). However, since the AI Act governs AI products placed on the market or put into service within the European Union, AI products developed within the EU but subsequently exported to third countries generally fall outside the Act's scope. This exclusion does not apply, however, if such AI products, once exported, subsequently trigger any of the conditions articulated in Article 2—for instance, if outputs generated by the AI system in a third country are later employed within the Union.

2. Asymmetric Regulation

One of the AI Act's most distinctive features is its asymmetric regulatory approach, which imposes stricter requirements on AI systems based on the level of risk they pose. The Act categorizes AI systems according to the potential risks they present to fundamental rights and health and safety. Certain AI practices that pose unacceptable risks are explicitly prohibited (Article 5). In contrast, AI systems deemed high-risk are subject to rigorous obligations before being placed on the market, including the implementation of risk assessment and mitigation strategies, the establishment of robust data governance frameworks, and adherence to technical documentation requirements (Section 2, AI Act).

Additionally, the Act places specific obligations on general-purpose AI systems that present systemic risks, such as undergoing model evaluations and implementing risk management systems (Article 55). Other AI systems, presumably posing a low risk—although the regulation does use this language and not explicitly acknowledge a lower risk—are subject to minimal regulatory requirements and may be freely placed on the EU market and used.

The asymmetric nature of the AI Act also extends to the regulatory responsibilities of different actors within the AI value chain. While AI providers are held to stringent requirements, deployers are subject to less burdensome obligations, such as ensuring that users receive appropriate instructions for safe use. Moreover, the AI Act differentiates between AI systems based on factors such as their intended use (e.g., research and development), computational power (e.g., large-scale general-purpose models), and accessibility (e.g., open-source systems). This tailored approach seeks to prevent overregulation while addressing the specific risks associated with different AI applications.

3. Innovative Version of Product-Safety Regime

The AI Act is designed as a product safety instrument, drawing inspiration from the regulatory framework for medical devices, which often involve software components.[10] Under the AI Act, high-risk AI systems must undergo a conformity assessment, be registered in an EU database, and bear the CE marking before being placed on the market. This approach enables the EU to leverage its long-established regulatory expertise in product safety, a field in which it has considerable institutional experience. However, a downside of this product-safety approach is that it imposes significant burdens on providers of AI systems. On the other hand, to support innovation, EU regulators have introduced AI regulatory sandboxes (Article 57). The AI Act mandates that Member States establish at least one AI regulatory sandbox at the national level. This provision aims to foster innovation within the EU, ensuring that the AI Act not only safeguards against risks but also promotes technological advancement.[11] In this regard, the AI Act represents a balanced approach between safety and innovation.

III. Regulatory Approach of the AI Act

1. A Risk-Based Approach?

The AI Act is generally described as a risk-based regulation, and it frames itself as such: "In order to introduce a proportionate and effective set of binding rules for AI systems, a clearly defined risk-based approach should be followed. That approach should tailor

[10] Almada & Petit, The EU AI Act: a medley of product safety and fundamental rights?, Robert Schuman Centre for Advanced Studies Research Paper 2023/59, 2023.

[11] Boura, The Digital Regulatory Framework through EU AI Act: The Regulatory Sandboxes' Approach, Athens JL 10, 2024, 385 ; Truby, et al., A sandbox approach to regulating high-risk artificial intelligence applications, European Journal of Risk Regulation 13.2, 2022, 270-294.

the type and content of such rules to the intensity and scope of the risks that AI systems can generate…" (Recital 26). Literature on the AI Act also classifies it as a risk-based regulation, given that it subjects AI systems to different requirements depending on the risks they pose.[12] The AI Act aims to create trustworthy AI that poses only acceptable risks.

There are three key components of the AI Act that support its classification as risk-based: the categorization of systems according to the risks they present, the concept of risk acceptability, and the emphasis on risk management. However, despite its original intention, the Act has evolved into a hybrid of risk-based and principles-based regulation, ultimately influenced by fundamental rights.

a) The Classification of AI systems

The classification of AI systems in the AI Act is not solely based on the risks they pose. Some categories do not directly refer to risk, which complicates the Act's characterization as risk-based. The six categories of AI systems are:

- Prohibited AI practices (e.g., social scoring, facial recognition, dark-pattern AI), which present unacceptable risks and are banned (Art. 5).
- High-risk AI systems, used in critical sectors like immigration, education, employment, and justice, that must undergo conformity assessments and be monitored to mitigate risks (Art. 6 and s.).
- AI systems that pose limited risks, such as chatbots, deepfakes, and emotion recognition, which require transparency (Art. 50).
- General-purpose AI models that do not pose risks, where providers must create technical documentation, establish copyright policies, and provide training content summaries (Art. 53).
- General-purpose AI models that pose systemic risks, such as those with unintended control issues or harmful consequences like discrimination or disinformation. These models must be notified to the Commission, undergo risk assessments and mitigation measures, and ensure adequate cybersecurity (Art. 55).
- AI systems with no risk, like spam filters and video games, which are not subject to requirements under the AI Act but may fall under other EU laws.

[12] See for example Fraser & Bello y Villarino, Where Residual Risks Reside: A Comparative Approach to Art 9(4) of the European Union's Proposed AI Regulation, SSRN, 2022 ; Mahler, Between Risk Management and Proportionality: The Risk-Based Approach in the EU's Artificial Intelligence Act Proposal, Nordic Yearbook of Law and Informatics 247, 2022 ; and Chamberlain, The Risk-Based Approach of the European Union's Proposed Artificial Intelligence Regulation: Some Comments from a Tort Law Perspective, European Journal of Risk Regulation, 2022.

Among these six categories, only high-risk AI systems and general-purpose AI models posing systemic risks are subject to stringent regulation due to the risks they pose. Conversely, other categories are either entirely prohibited—reflecting an ex ante determination by regulators that their risks are inherently unacceptable—or subjected to minimal compliance obligations, presumably due to their lower risk profile. Notably, general-purpose AI systems without systemic risk implications nonetheless fall under regulatory oversight in the AI Act. This inclusion underscores the observation that risk assessment alone does not fully account for the Act's regulatory architecture, highlighting the presence of additional, non-risk-based considerations in shaping its legislative framework.

Although the AI Act is frequently characterized by scholars and regulatory bodies as adopting a risk-based framework, a nuanced examination reveals complexities that challenge this prevalent interpretation. Specifically, the Act's classification of AI systems, while ostensibly risk-oriented, incorporates criteria extending beyond mere risk assessments. Consequently, characterizing the AI Act strictly as "risk-based" oversimplifies its conceptual and regulatory underpinnings and fails to fully encapsulate the multifaceted objectives and underlying rationale inherent in its design and future application.

b) The Notion of Risk Acceptability

The AI Act integrates a mix of risk-based, principles-based, and rights-based approaches. Specifically, the Act assesses the acceptability of risks through the prism of established principles for trustworthy artificial intelligence[13] and the imperative to safeguard fundamental rights. This integration inherently complicates any attempt to categorize the AI Act strictly within a conventional risk-based approach, instead reflecting a hybrid regulatory model where risk considerations are combined with principle-driven and rights-centered approaches.

aa) An Ethical-Principles-Based Regulation:

The AI Act blends a risk-based approach with principles-based regulation in an effort to promote trustworthy AI. The Act's objective is to "promote the uptake of human-centric and trustworthy artificial intelligence" (Art. 1). Trustworthy AI is central to discussions around AI regulation, and several policy frameworks aim to promote it. The AI Act combines the risk-based approach with principles-based elements to foster

[13] Johann, Wachter & Mittelstadt, Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk, Regulation & Governance 18.1, 2024, 3-32.

trustworthy AI: "While the risk-based approach is the basis for a proportionate and effective set of binding rules, it is important to recall the 2019 Ethics Guidelines for Trustworthy AI developed by the independent AI HLEG appointed by the Commission. 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 (...) The application of those principles should be translated, when possible, in the design and use of AI models." (Recital 27). These principles guide the application of AI regulation to ensure risks align with trustworthy AI principles.

The AI Act intertwined the risk-based regulatory approach with underlying ethical principles, thereby gauging the acceptability of AI-related risks through compliance with these normative benchmarks. As articulated by Laux et al., the European Commission adopts a constrained interpretation of "trustworthiness," predominantly equating it with the acceptability of AI risks[14]. Consequently, the Act's risk-based framework emphasizes adherence to ethical principles rather than merely adhering to conventional risk assessment methodologies. Illustratively, the regulatory requirements governing high-risk AI systems explicitly seek to achieve a "high level of trustworthiness" (Recital 64), signifying that the obligatory provisions are primarily designed to render associated risks acceptable within the broader conceptualization of trustworthy artificial intelligence.

bb) A Rights-Based Regulation:

The AI Act also incorporates a rights-based dimension, addressing risks to health, safety, and fundamental rights. One of the primary goals of the Act is to promote "human-centric" AI[15] that protects fundamental rights enshrined in the EU Charter, including democracy, the rule of law, and environmental protection (Art. 1). The Act references fundamental rights over one hundred times, and risk assessments are evaluated in light of their acceptability regarding the protection of these rights. Moreover, some AI practices are prohibited because they pose unacceptable risks to fundamental rights. As stated in Recital 28, AI systems used for manipulative, exploitative, and socially controlling practices are banned because they violate rights like non-discrimination, data protection, privacy, and children's rights. Moreover, AI systems are categorized as high-risk because they can potentially threaten fundamental rights.

[14] Laux, Wachter & Mittelstadt, Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk, Regulation & Governance 18.1, 2024, 3-32.

[15] Restrepo Amariles & Marcello Baquero, Promises and limits of law for a human-centric artificial intelligence, Computer Law & Security Review 48, 2023.

This fundamental rights dimension is highlighted by Almada and Petit, who argue that the AI Act is a blend of product safety regulation and fundamental rights protection[16]. Thus, providers of high-risk AI systems are required to assess their impact on fundamental rights and ensure compliance with risk management measures that protect these rights[17]. The regulation of AI systems is, therefore, inspired by rights-based instruments.

In conclusion, the AI Act combines risk-based, principles-based, and rights-based regulation, assessing all risks against the principles of trustworthy AI and the protection of fundamental rights.

c) The Management of Risks

Risk management plays a crucial role in AI regulation. Various standard-setting bodies have developed frameworks for AI risk management, such as the NIST AI Risk Management Framework (AI RMF 1.0)[18] and ISO/IEC 23894:2023 Information technology – Artificial Intelligence – Guidance on risk management[19]. Interestingly, the AI Act assigns a significant role to risk management, but only for high-risk AI systems and general-purpose AI models with systemic risks[20].

While EU regulators establish a broad category of high-risk AI systems based on their intended use and general characteristics, they delegate to providers the responsibility of identifying and mitigating the specific risks presented by individual systems through defined risk management requirements. Under Article 9 of the AI Act, high-risk AI systems must comply with specific regulatory obligations calibrated according to their identified risks. Providers of such systems must establish, implement, document, and maintain robust risk management systems tailored explicitly to these risks. Article 9 sets forth a structured framework for this risk management process: Articles 9(1) and 9(2) mandate the creation of a comprehensive risk management system; Articles 9(4) and 9(5) prescribe specific risk mitigation measures; and Articles 9(6) through 9(8) outline necessary testing procedures. Collectively, these provisions underscore the

[16] Almada & Petit, The EU AI Act: a medley of product safety and fundamental rights?, Robert Schuman Centre for Advanced Studies Research Paper 2023/59, 2023.

[17] Malgieri & Santos, Assessing the (severity of) impacts on fundamental rights, Computer Law and Security Review, 2025.

[18] NIST AI Risk Management Framework (AI RMF 1.0), January 2023, https://nvlpubs.nist.gov/nistpubs/ai/NIST.AI.100-1.pdf

[19] ISO/IEC 23894:2023, ISO/IEC 42001:2023, Information technology — Artificial intelligence — Management system, Ed. 1, 2023, https://www.iso.org/standard/81230.html

[20] Schuett, Risk management in the artificial intelligence act, European Journal of Risk Regulation 15.2 2024, 367-385.

regulatory principle that providers bear primary responsibility for recognizing and addressing the particular risks inherent in their high-risk AI systems.

Moreover, providers of general-purpose AI models that present systemic risks face additional risk management responsibilities compared to providers of other general-purpose AI systems. Specifically, Article 55 mandates that these providers rigorously evaluate their AI models using standardized assessment protocols, including adversarial testing. Providers must also identify, assess, and mitigate any systemic risks at the EU-wide level. Furthermore, they must document serious incidents and corrective actions taken in response to these events. Through these provisions, the AI Act places explicit emphasis on the accountability of providers to proactively assess, document, and address systemic risks inherent in their AI systems, underscoring a heightened regulatory scrutiny aimed at protecting EU-wide interests and stability.

The AI Act's emphasis on risk-based regulation does not imply an exclusive focus on risk management. Although central to the regulatory approach, risk management alone is insufficient to achieve the Act's broader objectives. The AI Act also incorporates principles-driven measures explicitly designed to ensure the development of trustworthy AI and to safeguard fundamental rights. Consequently, its scope extends beyond merely managing risks to promoting AI systems that are aligned with ethical standards and fundamental rights protections.

2. The Meta-Regulation Model Underlying the AI Act

a) Meta Regulation vs Conventional Regulation and Self-Regulation

The regulatory approach adopted by the AI Act provides a compelling illustration of meta-regulation in practice. Unlike traditional command-and-control regulatory frameworks, which prescribe precise outcomes and rigid directives to regulated entities[21], the AI Act refrains from specifying exact requirements that AI systems must fulfill. Instead, it places greater responsibility on providers and other stakeholders across the AI value chain, empowering them to devise internal processes designed to achieve regulatory compliance.

This approach notably diverges from conventional regulatory models, which typically minimize discretion by detailing explicit actions required of regulated parties. By contrast, the AI Act introduces a significant degree of flexibility, incentivizing actors to develop their own compliance frameworks. Nonetheless, the AI Act does not embody pure self-regulation, given its compulsory nature, oversight by a structured network of enforcement authorities, and substantial penalties for non-compliance. Consequently,

[21] Bardach & Kagan, Going by the Book: The Problem of Regulatory Unreasonableness, 1982, Philadelphia, PA: Temple University Press.

the AI Act occupies an intermediate position, bridging traditional prescriptive regulation and voluntary self-regulation.

Coglianese and Mendelson propose the concept of a "regulatory discretion pyramid" as a framework to categorize varying regulatory approaches[22]. As illustrated in Figure 1, meta-regulation occupies an intermediate position between conventional command-and-control regulation and pure self-regulation. Unlike conventional regulation, meta-regulation grants regulated entities considerable discretion, incentivizing them to develop internal regulatory systems tailored to meet broader compliance goals. This description closely aligns with the regulatory approach underlying the AI Act. For instance, providers of high-risk AI systems must independently design, implement, and maintain comprehensive risk management frameworks. Similarly, providers of general-purpose AI models are obligated to produce and regularly update detailed technical documentation and policies to ensure compliance with applicable EU laws. Furthermore, general-purpose AI systems that pose systemic risks require providers to establish internal compliance procedures, including rigorous performance testing, systemic risk assessment, and robust risk mitigation processes. In brief, stakeholders throughout the AI value chain are expected to assume significant responsibility for internal governance to ensure adherence to regulatory requirements.

Figure 1: Regulatory Discretion Pyramid

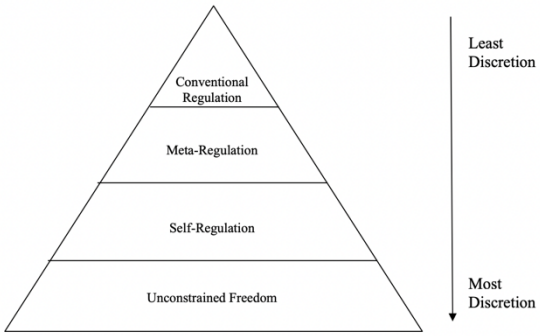


Fig. 1, Regulatory Discretion Pyramid, Coglianese, Cary and Mendelson, Evan, Meta-Regulation and Self-Regulation (2010).

[22] Coglianese & Mendelson, Meta-Regulation and Self-Regulation, The Oxford Handbook on Regulation, Cave, Baldwin, Lodge, eds., 2010.

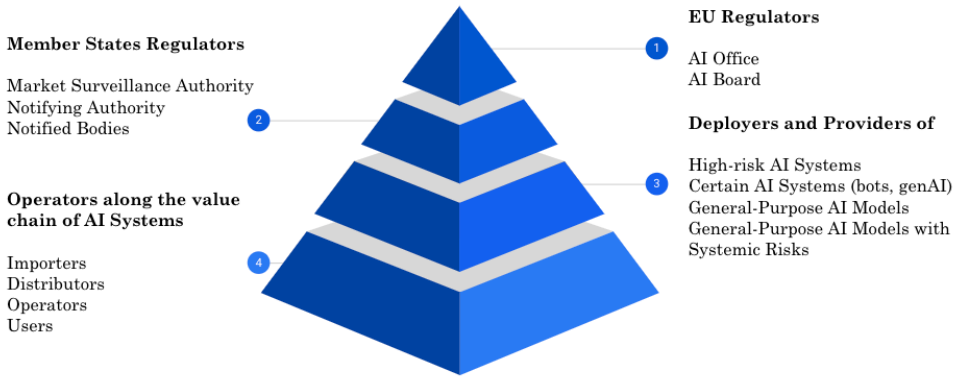
b) The Institutionally Distributed Regulation of the AI Act

Meta-regulation also entails efforts by governmental authorities to promote and oversee self-regulation[23]. This is precisely what the AI Act embodies, through the involvement of various regulatory intermediaries[24]. As noted by Tervo et al., regulatory intermediaries play a pivotal role in the implementation and enforcement of the AI Act[25]. Notably, the Act involves both actors with official mandates (e.g., the AI Office, notified bodies) and those without formal mandates (e.g., providers of AI systems), all of whom contribute to the Act's implementation, monitoring, and enforcement[26].

The Act establishes a tiered regulatory framework, involving multiple entities at different levels of governance:

- Supervisory roles: These are divided between the EU and Member States, with closer oversight provided by notified bodies for certain high-risk AI systems.
- Providers and deployers: These actors are responsible for ensuring that their AI systems comply with the AI Act and for developing their own internal regulations.
- Actors throughout the AI value chain: These actors also play a role in ensuring compliance and enforcing the AI Act's provisions.

This institutionally distributed regulation is in line with the meta-regulatory approach as the regulatory process is shared across various levels of authority and responsibility (see Fig. 2).



[23] Coglianese & Mendelson, Meta-Regulation and Self-Regulation, The Oxford Handbook on Regulation, Cave, Baldwin, Lodge, eds., 2010.

[24] Erkki, Väyrynen, & Iivari, Increasing understanding about the role of regulatory intermediaries in regulation–a scoping review and implications for the European AI Act, MCIS 2024 Proceedings, 2024.

[25] Ibid.

[26] Erkki, Väyrynen, & Iivari, Increasing understanding about the role of regulatory intermediaries in regulation–a scoping review and implications for the European AI Act, MCIS 2024 Proceedings, 2024.

Fig. 2, Institutionally distributed regulation of the AI Act.

c) The Co-Drafting of the Law in the AI Act

Another salient feature of the AI Act's meta-regulatory approach is its sophisticated institutional design facilitating collaborative rulemaking among diverse regulatory actors. This structured interplay ensures that the regulatory framework remains both robust and adaptive to technological innovation. Specifically, EU and Member State regulatory bodies assume primary responsibility for enacting and enforcing the AI Act and its associated national provisions. In addition, the European Commission provides critical interpretive guidance, delineating practical implementation details (Art. 6), and retains authority to adopt implementing acts that establish uniform specifications (Art. 41).

Moreover, standardization organizations play a pivotal role by developing technical standards at the Commission's request, thereby concretizing compliance expectations and enhancing predictability for stakeholders (Art. 40). Complementing these efforts, the AI Office orchestrates the drafting of Codes of Practice, integrating contributions from a broad spectrum of relevant stakeholders—ranging from civil society groups and academia to industry representatives, AI providers, and deployers—thus embedding a diverse array of expert insights into regulatory practice (Art. 56).

Finally, the providers and deployers of AI systems themselves actively participate in shaping regulatory compliance through the creation of internal policies and documentation aligned with the Act's obligations. This distributed governance structure exemplifies a dynamic and responsive regulatory model, effectively balancing the imperative for technological advancement with critical protections for fundamental rights, market integrity, and societal stability.

Gormet and Maxwell underscore the pivotal role that standards and codes of practice play within the EU AI Act's regulatory architecture[27]. These instruments frequently serve as vehicles for co-drafting, wherein technical experts, industry operators, and other relevant stakeholders collaboratively develop compliance frameworks[28]. For instance, the draft code of practice addressing general-purpose AI introduces specific

[27] Gornet & Maxwell, The European approach to regulating AI through technical standards, Internet Policy Review 13.3, 2024, 1-27

[28] Ibid.

obligations that reflect this cooperative regulatory paradigm[29]. This collaborative approach not only enhances stakeholder buy-in but also ensures that the regulatory framework remains adaptive and effective.

Overall, the meta-regulation regime of the AI Act reflects a highly collaborative and dynamic regulatory approach, where different stakeholders play an active role in shaping AI system regulation.

IV. Comparison between the AI Act, the DSA and the GDPR

Another distinctive feature of the AI Act lies in its comparative relationship with other European regulatory instruments, notably the General Data Protection Regulation (GDPR) and the Digital Services Act (DSA). All these frameworks share an underlying commitment to risk-based regulation; however, they significantly diverge in the specific regulatory regimes and compliance obligations they establish. While each embodies elements of meta-regulation and employs risk-based frameworks, critical variations are evident in their respective institutional architectures and implementation methods. Specifically, the AI Act uniquely operationalizes meta-regulation through an emphasis on collaborative processes involving standardization bodies and extensive stakeholder engagement, creating a more adaptive and responsive regulatory landscape.

1. Risk-Based Approach and Meta-Regulation: Commonalities

A notable characteristic of the AI Act, and indeed of related European regulations such as the GDPR and the DSA, is their adoption of a hybrid regulatory model blending risk-based and conventional regulatory approaches. Although all three frameworks employ risk-based elements to safeguard fundamental rights, public health, and safety, none rely exclusively on this model. Instead, they differentiate regulatory obligations based on the perceived level of risk: certain clearly defined activities or technologies attract prescriptive minimum standards, while higher-risk or more sensitive areas necessitate comprehensive internal risk management processes embedded deeply into compliance programs.

[29] Amit, The Regulation of GPAI Model Providers Under the EU AI Act, Regulatory Competition in the Digital Economy: Artificial Intelligence, Data, and Platforms. Cham: Springer Nature Switzerland, 2025. 119-137; Riede, Borelli, Kirchmair. EU AI Act Unpacked #19: General-Purpose AI Code of Practice – An overview, Freshfield Technology Quotient, 09 December 2024, https://technologyquotient.freshfields.com/post/102jqlt/eu-ai-act-unpacked-19-general-purpose-ai-code-of-practice-an-overview.

For instance, under the GDPR, large-scale data processing activities mandate structured risk assessments and corresponding mitigations.[30] Similarly, the AI Act mandates robust risk management protocols specifically for high-risk AI systems, and the DSA targets systemic risks posed by very large online platforms (VLOPs). This dual strategy—conventional regulation combined with meta-regulatory methods—enables the EU to address known digital threats explicitly while maintaining the flexibility required to respond effectively to emerging risks through collaborative, self-regulatory mechanisms involving relevant stakeholders[31].

Furthermore, the GDPR, DSA, and AI Act collectively exemplify a meta-regulatory approach by mandating that regulated entities establish and administer internal compliance frameworks. Rather than merely prescribing specific compliance obligations, these regulatory instruments compel firms to build and sustain internal compliance mechanisms, including dedicated personnel, risk assessment procedures, and internal oversight structures[32]. Such an approach underscores a preference for leveraging internal governance capacities, particularly in addressing intricate or evolving risks through adaptive self-regulation.

Moreover, these regulations adopt an institutionally decentralized regulatory model. This framework encompasses a network of European and Member State regulatory bodies and actively involves diverse stakeholders—governmental institutions, private sector actors, civil society groups, and academia—in the development, monitoring, and enforcement processes. Consequently, this pluralistic governance structure enhances regulatory responsiveness, adaptability, and legitimacy by ensuring broad participation and expertise in the management of digital risk landscapes.

2. Variations in the Meta-Regulation Approach

Despite their shared reliance on meta-regulation, the GDPR, DSA, and AI Act each adopt distinct approaches to internal compliance, reflective of their unique regulatory objectives and targeted areas of concern[33]. Under the GDPR, data controllers are required to establish suitable technical and organizational measures for safeguarding personal data (Article 24). While a dedicated compliance role is not universally mandated,

[30] Binns, Data protection impact assessments: a meta-regulatory approach. International Data Privacy Law. 2017, 22-35.

[31] See for example, on the GDPR: Binns, Data protection impact assessments: a meta-regulatory approach. International Data Privacy Law. 2017, 22-35.

[32] Jovanić, Platform Regulation: Ex Ante Tools and Strategies in the Regulatory Toolkit, Repositioning Platforms in Digital Market Law, 2024, 1-34.

[33] Pedro Rubim Borges & Restrepo Amariles, Law-jobs in the algorithmic society, International Journal of Law in Context 19.1, 2023, 1-12.

entities handling sensitive information or engaging in large-scale data processing must appoint a Data Protection Officer (DPO) (Article 37). The DPO oversees compliance, liaises with supervisory authorities, and serves as the primary contact for data subjects. Notably, the GDPR emphasizes the independence and adequate resourcing of the DPO, protecting this function from undue interference or penalization, thereby embedding data protection responsibilities at the highest organizational levels. Additionally, the GDPR introduces voluntary certification mechanisms (Article 42), offering regulated entities an avenue to demonstrate compliance proactively through third-party certification bodies. This voluntary dimension underscores a self-regulatory flexibility distinct from the more prescriptive requirements found in other EU digital regulations

The DSA establishes detailed compliance requirements primarily directed toward digital platforms, particularly VLOPs. Although the regulation does not universally mandate general monitoring obligations for intermediary service providers, it explicitly requires platforms to appoint qualified personnel responsible for specific content moderation functions (Article 20.6). For VLOPs, the compliance regime is significantly more rigorous. These platforms must undergo independent annual audits assessing their adherence to the DSA's provisions (Article 37), incurring the associated costs themselves. Additionally, VLOPs are obligated to maintain a distinct, adequately resourced compliance function led by compliance officers empowered to oversee platform operations and ensure adherence to regulatory requirements (Article 41). This compliance function must actively coordinate with the highest management levels, underscoring an organizational commitment to comprehensive risk management and regulatory compliance.

In contrast, the AI Act provides a more flexible framework for internal compliance compared to both the GDPR and DSA. It does not universally mandate the establishment of dedicated compliance departments or positions for all AI system providers. Instead, for high-risk AI systems, providers must implement robust quality management systems designed to embed compliance throughout the AI system's lifecycle, including procedures for design, testing, evaluation, and data governance (Article 17). Additionally, high-risk AI systems must undergo conformity assessments, which may be executed internally or through externally accredited notified bodies (Annex VI and VII). Furthermore, deployers of high-risk AI systems are required to ensure effective human oversight, structured flexibly according to the specific operational context, provided the individuals involved are sufficiently trained and resourced (Article 26).

This comparative analysis highlights the distinctiveness of the AI Act's meta-regulatory approach, reflecting its product-safety orientation. Unlike the GDPR and DSA, which primarily regulate providers or service operators, the AI Act places a particular emphasis on regulating AI systems themselves, thereby offering regulated entities greater flexibility in structuring internal compliance mechanisms.

V. Conclusion

In this analysis, we have examined the regulatory approach of the AI Act, highlighting its distinctive integration of risk-based, principle-based, and rights-based methodologies. Central to the AI Act is its meta-regulatory framework, which significantly influences both its implementation and enforcement, yielding notable advantages as well as meaningful challenges.

The AI Act's regulatory structure confers several important advantages. Foremost among these is the alleviation of regulatory burdens on public authorities, as compliance responsibilities are substantially delegated to industry stakeholders. This delegation allows regulators to prioritize broader strategic oversight rather than engage in granular, direct enforcement activities. Additionally, by leveraging self-regulation, the Act enables regulators to capitalize on the specialized expertise of AI providers and other industry participants, facilitating more effective and knowledgeable compliance practices. The inherent flexibility of the Act's meta-regulatory approach further supports innovation, fostering regulatory adaptability conducive to technological advancement without sacrificing essential oversight. The collaborative nature of compliance, distributed across multiple actors—including providers, deployers, and intermediaries—enhances the robustness and effectiveness of monitoring and enforcement mechanisms.

Despite these benefits, the meta-regulatory approach adopted by the AI Act is not without its drawbacks. Its effectiveness significantly depends on the voluntary cooperation and consistent engagement of regulated entities, posing risks if compliance proves sporadic or incomplete. There remains the possibility of misinterpretation or strategic evasion of the Act's provisions, potentially compromising regulatory objectives. Additionally, the complexity and extensive requirements of the Act impose substantial compliance burdens on industry actors, particularly concerning rigorous risk assessment obligations and the establishment of comprehensive internal compliance systems. Given that risk assessments, especially those involving fundamental rights, are predominantly entrusted to regulated entities themselves, inconsistencies or inadequacies in mitigating risks could arise. Finally, the extensive reliance on external stakeholders for compliance raises legitimate concerns regarding the capacity of regulatory bodies to cultivate sufficient internal AI expertise, potentially limiting their ability to effectively oversee AI-related regulatory issues.

In conclusion, the AI Act offers a promising regulatory framework characterized by flexibility, innovation-friendly principles, and robust stakeholder collaboration. Nevertheless, the success of this regulatory model ultimately hinges on effective cooperation between regulators and industry stakeholders, the rigorous adoption and adherence to

regulatory norms by the latter, and the continuous development of regulatory capacity to address the dynamic complexities of artificial intelligence.

Chapter 4: Brussels Effect? The outside perspective

The US Perspective

Margaret Hu[1]

I. Introduction

The Artificial Intelligence (AI) revolution is philosophical as much as it is technological.[2] AI governance in a constitutional democracy demands an inquiry into how best to responsibly sustain the economic growth and impact of AI in a way that is consistent with democratic commitments: the separation of powers and rights-based guarantees.[3] Specifically, the impact of AI on national security and fundamental rights can be better visualized by interrogating the possibility of how the AI revolution is poised to offer a parallel or competitive form of both digital sovereignty and digital citizenship, and "democratic" governance.[4] When seen as a power and knowledge structure that may be in competition with the Age of Enlightenment, the Age of AI has the potential to displace both law and evidence-based reasoning with AI reasoning.

Because this AI-driven challenge to power ordering can be presented as consistent with a constitutional system of democratic republicanism—purportedly expanding efficient and accurate governing, and equality in knowledge access and digital economy participation on an unprecedented scale—its reported democratic legitimacy is elevated. The opacity of both the AI systems and obfuscation of its threats, however,

---

[1] William & Mary Law School.

[2] Hu, Margaret. Senate Testimony. U.S. Senate Homeland Security and Governmental Affairs Committee Hearing, Washington, D.C. (November 8, 2023), Senate Committee Hearing: "The Philosophy of AI: Learning from History, Shaping Our Future." See also Vallor, Shannon. The AI Mirror: How to Reclaim Our Humanity in the Age of Machine Thinking (2024).

[3] Hu, supra note 1. See also Hu, Margaret. AI Law and Policy (2025); Acemoglu, Daron, and Johnson, Simon. Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity (2023).

[4] Hu, supra note 1. See also Zuboff, Shoshana. The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power (2019); Cohen, Julie. Between Truth and Power: The Legal Constructions of Informational Capitalism (2019); Varoufakis, Yanis. Technofeudalism: What Killed Capitalism (2024).

means that patrolling democratic boundaries has been increasingly complex. The complexity of negotiating national security and fundamental rights is exacerbated by a redefining of what is a national security interest and what is a fundamental rights interest in the AI economy. Consequently, it is increasingly unclear how to craft regulatory responses that are legitimately protective of national security objectives and the fundamental rights of the polity.

II. Fundamental Rights in the AI Cool War

Understanding AI innovation policy in the United States (U.S.) demands more clearly anticipating how the economic ecosystem will be defined by the AI Cold War[5] that will run parallel to historical examples like the space race and nuclear arms race of the Cold War.[6] National security strategies in the AI economy, particularly in a new great power competition and the emergence of "digital empires",[7] are impacting fundamental rights and vice versa. Building a robust, inclusive AI economy and effective AI regulation requires acknowledging that the AI revolution involves protecting national security objectives, but not at the cost of rights and liberties. The AI regulation objectives of the U.S., EU, China, and other nations also must not be introduced for economic competition reasons alone for the global AI economy to thrive.

The AI economy not only forces the U.S. and other nations to compete technologically, but also legally and economically.[8] AI economic competition also situates the U.S., EU, China, Russia, and other digital empires militarily and, obviously, geopolitically. In the U.S. and many other nations, AI innovation is currently driven by technology and AI companies. This forces digital empires to either join or control and capture the private sector to assist sovereign nations in maintaining or remaking the international order. Consequently, the AI Great Power Competition is not one that democratic governments can undertake as a public endeavor alone; instead, it is one that is techno-economically centered. Military use of AI is expanding into realms of advanced weapons design, surveillance, as well as new areas beyond conventional physical force and into information warfare and information pollution, and AI systems will be central

[5] Takach, George S. Cold War 2.0: Artificial Intelligence in the New War between China, Russia, and America (2024).

[6] Feldman, Noah. Cool War: The United States, China, and the Future of Global Competition (2015); Sokolski, Henry. *China, Russia, and the Coming Cool War*, Nonproliferation Policy Education Center (June 2024). *See also* Rothkopf, David. *The Cool War: Cold War technology made war unthinkable[;] Cool War technology makes it irresistible*, Foreign Pol'y (Feb. 20, 2013).

[7] Bradford, Anu. Digital Empires: The Global Battle to Regulate Technology (2023).

[8] Ding, Jeffrey. Technology and the Rise of Great Powers: How Diffusion Shapes Economic Competition (2024).

to components of peer warfare.[9] This potential for cooperation and overcooperation between the private and public sectors to address real and perceived sovereign threats through private AI innovation risks undermining both national security and fundamental rights.[10]

Within this landscape of cyberwar and information warfare, EU, China, and other nations are now outpacing the U.S. in AI legal innovation and newly adopted frameworks for AI governance.[11] In addition to the *EU AI Act* and the *General Data Protection Regulation*, the EU has implemented other regulatory frameworks in the AI economy, such as the *Digital Markets Act*, the *Digital Services Act*, and other laws. Likewise, China has recently proposed a comprehensive AI regulatory framework through a set of AI and data regulations: the *Administrative Provisions on Algorithm Recommendation for Internet Information Services*; *Provisions on Management of Deep Synthesis in Internet Information Service*; the *Provisional Provisions on Management of Generative Artificial Intelligence Services*; *Trial Measures for Ethical Review of Science and Technology Activities*; and other measures.

Immediately after ChatGPT was released, China blocked the download of the app[12] and explained that the generative AI technology posed a significant threat to their national security.[13] "This is signal that they understood modern warfare in a way that US/EU perhaps is not conceptually grasping--AI warfare is different."[14] China soon thereafter introduced its own generative AI system that it controls,[15] and soon moved

[9] King, Anthony. "Digital targeting: Artificial intelligence, data, and military intelligence. Journal of Global Studies 9, no. 2 (2024): ogae009; Hunter, Lance Y., Craig D. Albert, Josh Rutland, Kristen Topping, and Christopher Hennigan. "Artificial intelligence and information warfare in major power states: how the US, China, and Russia are using artificial intelligence in their information warfare and influence operations." Defense & Security Analysis 40, no. 2 (2024): 239-269.

[10] See, e.g., Hu, supra note 1; Cohen, supra note 3; Balkin, Jack. "Free Speech is a Triangle." Columbia Law Review Vol. 118, Iss. 7 (2018): 2011-2055.

[11] Chatterjee, Mohar. Senate Intelligence Chair: China Leads the World on AI Rule, Politico (June 15, 2023), https://www.politico.com/news/2023/06/15/senate-intelligence-chair-china-leads-the-world-on-ai-rules-00102168. See also White House. National Security Strategy 8 (Oct. 2022). https://bidenwhitehouse.archives.gov/wp-content/uploads/2022/11/8-November-Combined-PDF-for-Upload.pdf.

[12] Ray, Siladitya. ChatGPT Reportedly Blocked on Chinese Social Media Apps, Forbes (Feb. 22, 2023), https://www.forbes.com/sites/siladityaray/2023/02/22/chatgpt-reportedly-blocked-on-chinese-social-media-apps-as-beijing-claims-ai-is-used-to-spread-propaganda/?sh=50eed661372c.

[13] AP, China Warns of Artificial Intelligence Risks (May 31, 2023), https://www.pbs.org/newshour/world/china-warns-of-artificial-intelligence-risks-calls-for-increased-national-security-measures.

[14] Hu, Margaret; Behar, Eliott; Ottenheimer, Davi. National Security and Federalizing Data Privacy Infrastructure for AI Governance, Fordham Law Review (2024).

[15] Cheng, Evelyn. *China's AI Chatbots Haven't Yet Reached the Public Like ChatGPT Did*, CNBC (Apr. 28, 2023), https://www.cnbc.com/2023/04/28/how-chinas-chatgpt-ai-alternatives-are-doing.html.

immediately to create a legal framework for regulating generative AI.[16] Former Chair of U.S. Senate Select Committee on Intelligence Foreign, Senator Mark Warner (D-Virginia), has explained that a joint legal and technical innovation strategy by China and other nations may give other sovereign powers a competitive strategic advantage in national security.[17] The U.S. has taken measures to stymie Chinese AI development with export controls on AI model weights and advanced GPUs, but the competition is also a race, with the U.S. seeking to develop and maintain a lead in AI technologies. To that end, AI regulations and safeguards have carved out exceptions for intelligence activities and classified systems from oversight in sectors like law enforcement, national security, and other national security systems. Sectors that will be controlling and observing surveillance systems, speech processing, and data collection.[18]

The U.S. has fallen behind in regulating AI. The rapid advancement of AI and promotion of AI development is starting to run at odds with fundamental rights of free speech and privacy, and due process and equal protection, with the allegation that AI regulation is not only inconsistent with national security aims of the U.S., but also the expressive freedom guarantees of the First Amendment of the Bill of Rights of the U.S. Constitution.[19]

III. Conclusion

The AI revolution is distinct from past industrial revolutions in that it is more akin to the Age of Enlightenment than the Age of the Internet. How and why this distinction matters demands novel legal and national security strategies, particularly in AI regulation and fundamental rights-based governance developments, and in underscoring data privacy, and information security and integrity. It also necessitates interdisciplinary pedagogical innovation, AI literacy and upskilling, and integrating AI rights-based and risk-based awareness in global workforce preparedness. The increasingly complex nature of global economic dynamics and power conflicts necessitates integrating and

[16] Yang, Zeyi. *Four Things to Know About China's New AI Rules in 2024*, MIT TECHNOLOGY REV. (Jan. 17, 2024), https://www.technologyreview.com/2024/01/17/1086704/china-ai-regulation-changes-2024/.

[17] Chatterjee, *supra* note 8. *See also* Rajtmajer, Sarah, and Susser, Daniel. *Automated Influence and the Challenge of Cognitive Security*, Association for Computing Machinery, ACM ISBN 978-1-4503-7561-0/20/04 (2020), https://philpapers.org/archive/RAJAIA-2.pdf; Sheehan, Matt. *China's AI Regulations and How They Get Made* , Carnegie Endowment (July 10, 2023), https://carnegieendowment.org/2023/07/10/china-s-ai-regulations-and-how-they-get-made-pub-90117.

[18] Patel, Faiza. *Unregulated AI in national security poses risks to privacy, civil rights, and public trust*. AI Insight Forum: National Security (Dec. 6, 2024), https://www.brennancenter.org/our-work/research-reports/statement-faiza-patel-ai-insight-forum-national-security.

[19] Balkin, Jack. "Free Speech Versus the First Amendment." UCLA Law Review, Vol. 70 (2023): 1206-1273.

expanding knowledge on reinforcing democratic institutions and cognitive security, national security, and fundamental rights within emerging AI policy frameworks.

AI law and policy impacts sovereignty and national security, and a range of fundamental rights, including expressive freedoms, data and information privacy and protection, self-determination and autonomy, and cognitive security. No amount of investment in workforce AI upskilling and AI infrastructure investment will succeed unless both policymakers and the polity understand this necessary cooperation or cooperation between the private and public sectors in the AI cool war. Further, unless further inquiry is trained on how national security and fundamental rights protections might be undermining the rule of law and the social contract, proposed AI regulation in the U.S., EU, and other democratic nations might be working at cross purposes. Ultimately, the negotiating a balance between national security interests and fundamental rights interests in an AI economy may undermine both.

AI Governance and Asia Aspect

I-Ping Wang[1]

I. Introduction

In recent years, AI technology has rapidly advanced and been widely applied across various fields, such as facial recognition, targeted advertising, biomedical technology, and text and image generation. However, while AI applications bring convenience, they also raise concerns about potential risks. For instance, when AI is used to evaluate personal credit or job performance, it may introduce biases, discrimination, or inaccuracies, leading to erroneous assessments or even incorrect decisions. Therefore, guiding AI development to prevent harm is an issue that warrants close attention.

This article examines the regulatory approaches of China, Japan, Singapore, and Taiwan in order of decreasing regulatory intensity. By doing so, it explores how governments manage AI and highlights key regulatory considerations. However, given the vast scope of AI-related regulations and the constraints of time and space, this article focuses only on regulations relevant to the discussion.

II. China

1. Current Regulations

In China there are two effective regulations, namely the "Interim Measures for the Administration of Generative Artificial Intelligence announced Services" and the "Provisions on the Administration of Deep Synthesis of Internet-Based Information Services". On September 9, at the main forum of the 2024 National Cybersecurity Awareness Week, the National Cybersecurity Standardization Technical Committee released the "Artificial Intelligence Security Governance Framework 1.0".

[1] Professor of National Taipei University, wangyiping@mail.ntpu.edu.tw.

a) Interim Measures for the Administration of Generative Artificial Intelligence Services[2]

There was a similar process for remote real-time biometric identification systems, as once again, the European Parliament proposed—unsuccessfully—prohibiting all use of such systems without exception in all settings (amendment 220), not only those related to law enforcement. The same was true for Parliament's proposal to also prohibit 'AI systems for the analysis of recorded footage of publicly accessible spaces through 'post' remote biometric identification systems, unless they are subject to a pre-judicial authorisation in accordance with Union law and strictly necessary for the targeted search connected to a specific serious criminal offense as defined in Article 83 (1) of TFEU that already took place for the purpose of law enforcement', which also ultimately failed to be adopted (amendment 227).

aa) Introduction

This regulation has been in effect since August 15, 2023. According to Article 4 providers and users of generative AI services must comply with the following requirements:

(1) Uphold the core socialist values and refrain from generating content that incites subversion of state power, seeks to overthrow the socialist system, endangers national security and interests, damages the national image, incites secession or undermines national unity and social stability, promotes terrorism or extremism, propagates ethnic hatred or discrimination, contains violence, pornography, or false and harmful information, or any other content prohibited by laws and administrative regulations.
(2) Take effective measures during algorithm design, training data selection, model generation and optimization, and service provision to prevent discrimination based on ethnicity, religion, nationality, region, gender, age, profession, health status, and other factors.
(3) Respect intellectual property rights and business ethics, protect trade secrets, and refrain from using advantages in algorithms, data, or platforms to engage in monopolistic or unfair competition practices.
(4) Respect the legitimate rights and interests of others, avoid harming others' physical or mental well-being, and refrain from infringing upon others' portrait rights, reputation rights, honor rights, privacy rights, or personal information rights.
(5) Based on the characteristics of the service type, adopt effective measures to enhance the transparency of generative AI services and improve the accuracy and reliability of generated content.

[2] Interim Measures for the Administration of Generative Artificial Intelligence Services, 02. January 2025, https://www.cac.gov.cn/2023-07/13/c_1690898327029107.htm.

bb) Technological Development and Governance

Articles 5 to 8 set forth provisions regarding technological development and governance, specifically regulating the data collection, training processes, and other activities of generative AI service providers (hereinafter referred to as “providers”). The specific obligations are detailed in Articles 7 and 8.

Article 7 requires providers to conduct pre-training, optimization training, and other training data processing activities in accordance with the law and to comply with the following requirements:

(1) Use data and foundational models with legitimate sources.
(2) Ensure that intellectual property rights are not infringed when involving copyrighted content.
(3) Obtain individual consent or meet other conditions stipulated by laws and administrative regulations when processing personal information.
(4) Take effective measures to improve the quality of training data, enhancing its authenticity, accuracy, objectivity, and diversity.
(5) Comply with other relevant provisions of the Cybersecurity Law of the People's Republic of China, the Data Security Law of the People's Republic of China, the Personal Information Protection Law of the People's Republic of China, and related regulatory requirements issued by competent authorities.

Article 8 stipulates that when conducting data annotation in the research and development of generative AI technology, providers must establish clear, specific, and operational annotation guidelines in accordance with this regulation. They must also conduct quality assessments of annotated data, perform sample verification to ensure annotation accuracy, provide necessary training to annotation personnel to enhance their legal awareness, and supervise and guide them to ensure compliance with standardized annotation practices.

cc) Service Standard

Articles 9 to 15 outline the regulations regarding service standards. Article 9 stipulates that providers must assume legal responsibility as network information content producers and fulfill their cybersecurity obligations. When handling personal information, they must comply with legal requirements as personal information processors and fulfill personal information protection obligations. Providers must sign service agreements with users of their generative AI services (hereinafter referred to as "users") to clearly define the rights and obligations of both parties.

Article 10 requires providers to clearly define and publicly disclose the target audience, applicable scenarios, and intended uses of their services. They must guide users in

scientifically and rationally understanding and lawfully using generative AI technology while taking effective measures to prevent minors from becoming overly dependent on or addicted to generative AI services.

Article 11 mandates that providers have a legal obligation to protect users' input data and usage records. They must not collect unnecessary personal information, unlawfully retain input data and usage records that can identify users, or illegally share such information with third parties. Providers must also promptly process users' requests regarding access, copying, correction, supplementation, and deletion of their personal information in accordance with the law.

Article 12 requires providers to mark generated content such as images and videos in accordance with the Provisions on the Administration of Deep Synthesis in Internet Information Services. Article 13 mandates that providers ensure safe, stable, and continuous services during the provision of their services, guaranteeing normal user access.

Article 14 requires providers to take immediate action upon discovering illegal content, including halting its generation or transmission, removing it, implementing corrective measures such as model optimization training, and reporting the issue to the relevant authorities. If a provider discovers that a user is using generative AI services for illegal activities, they must take appropriate actions per legal and contractual requirements, such as issuing warnings, restricting functionalities, suspending, or terminating services. They must also retain relevant records and report the issue to the competent authorities.

Article 15 requires providers to establish and improve complaint and reporting mechanisms. They must set up convenient channels for complaints and reports, publicly disclose processing procedures and response timelines, and promptly handle and respond to public complaints and reports.

b) Provisions on the Administration of Deep Synthesis of Internet-Based Information Services[3]

This regulation has been in effect since January 10, 2023.

aa) General Provisions

Article 6 stipulates that no organization or individual shall use deep synthesis services to create, copy, publish, or disseminate information prohibited by laws and administrative regulations. Deep synthesis services must not be used for activities that endanger national security and interests, damage the national image, harm public interests, disrupt economic and social order, or infringe upon the legitimate rights and interests of

[3] Provisions on the Administration of Deep Synthesis of Internet-Based Information Services, 02. January 2025, https://www.gov.cn/zhengce/zhengceku/2022-12/12/content_5731431.htm.

others. Providers and users of deep synthesis services are prohibited from using these services to create, copy, publish, or disseminate false news information. When reposting news content generated through deep synthesis services, the source must be a legally authorized internet news provider.

Article 7 stipulates that deep synthesis service providers must assume primary responsibility for information security and establish comprehensive management systems, including user registration, algorithm and mechanism reviews, technology ethics reviews, content review mechanisms, data security, personal information protection, anti-telecom fraud measures, and emergency response protocols. They must also implement secure and controllable technical safeguards.

Article 8 stipulates that deep synthesis service providers must establish and publicly disclose management rules and platform guidelines, improve service agreements, and fulfill their management responsibilities according to laws and agreements. They must prominently inform both technical supporters and users of their information security obligations.

Article 9 stipulates that deep synthesis service providers must verify the real identity of users through mobile phone numbers, identification documents, unified social credit codes, or the national online identity authentication public service system. Providers must not offer information publishing services to users who have not completed real identity verification.

Article 10 stipulates that deep synthesis service providers must strengthen content management by employing technical or manual review methods to assess user input and synthetic content. Providers must establish and maintain a database for identifying illegal and harmful content, improving database standards, rules, and procedures while retaining relevant network logs. If illegal or harmful content is discovered, providers must take appropriate action, retain relevant records, and promptly report to the cybersecurity and other relevant authorities. They must also take necessary actions against offending users, including issuing warnings, restricting functionalities, suspending services, or closing accounts.

Article 11 stipulates that deep synthesis service providers must establish a rumor-correction mechanism. If false information is discovered being created, copied, published, or disseminated using deep synthesis services, providers must take immediate corrective measures, retain relevant records, and report the issue to the cybersecurity and relevant authorities.

Article 12 stipulates that deep synthesis service providers must provide convenient channels for user appeals, public complaints, and reports. They must disclose their handling procedures and response timelines and ensure timely processing and feedback on complaints and reports.

Article 13 stipulates that application distribution platforms, such as internet app stores, must fulfill their security management responsibilities, including review before listing, daily supervision, and emergency response. They must verify whether deep synthesis applications comply with security assessments and filing requirements. If an application violates national regulations, the platform must take appropriate measures, such as refusing to list it, issuing warnings, suspending services, or removing it from the platform.

bb) Data and Technology Management Regulations

Article 14 requires that deep synthesis service providers and technology supporters must strengthen the management of training data and take necessary measures to ensure data security. If training data contains personal information, it must comply with relevant personal information protection regulations. If deep synthesis service providers and technology supporters offer biometric information editing functions (such as facial or voice data), they must inform users that they are legally required to notify the individuals whose biometric data is being edited and obtain their explicit consent.

Article 15 stipulates that deep synthesis service providers and technology supporters must enhance technical management by regularly reviewing, evaluating, and verifying the mechanisms of synthetic generation algorithms. If they provide models, templates, or other tools with the following capabilities, they must conduct security assessments either independently or through professional institutions:

(1) Generating or editing biometric information such as facial or voice data.
(2) Generating or editing non-biometric information, such as special objects or scenes, that may involve national security, national image, national interests, or public interests.

Article 16 requires that deep synthesis service providers must apply technical measures to add identifiers to generated or edited content that do not interfere with user experience. They must also retain log information in accordance with laws, administrative regulations, and national regulations.

Article 17 stipulates that deep synthesis service providers must prominently label generated or edited content in an appropriate location or area when providing deep synthesis services that could cause public confusion or misidentification, specifically for:

(1) Intelligent dialogue, intelligent writing, or other text-generation or editing services that simulate human interactions.
(2) Speech synthesis, voice imitation, or other services that generate or significantly alter an individual's voice characteristics.
(3) Facial generation, facial replacement, facial manipulation, posture control, or other services that generate or significantly alter an individual's image in pictures or videos.

(4) Immersive virtual reality or other simulated scene generation or editing services.
(5) Other services that generate or significantly alter information content. For deep synthesis services not explicitly listed above, providers must offer a prominent identification function and notify users that they have the option to apply visible identifiers.

Article 18 stipulates that no organization or individual shall use technical means to delete, tamper with, or conceal the deep synthesis identifiers required by Articles 16 and 17 of these provisions.

c) Artificial Intelligence Security Governance Framework 1.0

The framework establishes principles for AI security governance, including inclusiveness and prudence to ensure security, risk-oriented and agile governance, a combination of technology and management for coordinated response, and open cooperation for co-governance and shared benefits. Following the concept of risk management and closely integrating the characteristics of AI technology, it analyzes the sources and manifestations of AI risks. It addresses inherent security risks such as model algorithm security, data security, and system security, as well as application security risks across network, physical, cognitive, and ethical domains. The Framework proposes corresponding technical responses, comprehensive prevention measures, and guidelines for the secure development and application of AI.

2. In Discussion

There is a draft AI Act in discussion. I can see it in the Legislative Schedule of the State Council for 2023, 2024, but there is no relevant information about the regulations available online.

3. Conclusion

In 2023, China issued two regulations on generative artificial intelligence and deep synthesis in internet information services, both of which carry legal effect and impose responsibilities on relevant stakeholders. Articles 9 to 15 of the "Interim Measures for the Management of Generative Artificial Intelligence Services" impose multiple obligations on service providers. These include implementing effective measures to prevent minors from excessive reliance on or addiction to generative AI services, refraining from collecting unnecessary personal information, prohibiting the illegal retention of user-identifiable input data and usage records, and marking AI-generated content such as images and videos in accordance with regulations. However, certain provisions, such as

Article 9, which requires service providers and users to sign a service agreement clarifying their rights and obligations, and Article 13, which mandates providers to offer safe, stable, and continuous services to ensure normal user access, do not specifically address AI-related risks but rather constitute general contractual obligations. As a result, these regulations appear to encompass all issues related to generative AI rather than focusing solely on risks unique to the technology. Whether this regulatory approach is appropriate remains a subject of discussion.

The "Provisions on the Administration of Deep Synthesis of Internet-Based Information Services" specifically address the provision of internet information services using deep synthesis technology and have a broader scope of regulation. In the general provisions, the regulations prohibit any organization or individual from using deep synthesis services for illegal activities and impose multiple obligations on service providers. These include assuming primary responsibility for information security, verifying the real identities of users, reviewing user-inputted data and synthesized content through technical or manual means, and establishing a comprehensive misinformation correction mechanism. However, Article 8, which requires deep synthesis service providers to formulate and publicly disclose management rules, platform guidelines, and improve service agreements, falls again under general contractual obligations rather than addressing risks unique to AI. Regarding data and technology management, the regulations impose specific requirements on deep synthesis service providers and technical support entities, such as implementing necessary measures to ensure the security of training data and strengthening technical management. Additionally, Article 17 (1), lists specific scenarios where deep synthesis service providers must provide prominent public disclosures if the synthesized content may cause public confusion or misidentification. For other types of deep synthesis content, providers are only required to offer a prominent labeling function and notify users that they can apply such labels.

However, following the enactment of the two legally binding regulations mentioned above, China released the AI Governance Framework in 2024. Based on the concept of risk management, this framework proposes corresponding technical responses and comprehensive preventive measures, as well as guidelines for the safe development and application of AI. Since the framework itself is not legally binding, its purpose and its relationship with the aforementioned regulations are noteworthy considerations.

III. Japan

1. Current: AI Guidelines for Business Version 1.0

a) Introduction

Japan currently has no law specifically directed to regulating AI at this time. On April 19, 2024 the Japanese government published new "AI Guidelines for Business Version 1.0"[4] (hereinafter referred to as the "Guidelines") which consolidated and replaced three previously existing guidelines. The Guidelines states that three values should be respected as basic philosophies: dignity, diversity and inclusion, sustainability.[5] Based on the basic philosophies the guidelines proposes ten common guiding principles: human-centric, safety, fairness, privacy protection, ensuring security, transparency, accountability, education/literacy, ensuring fair competition, and innovation.[6] The Guidelines are principles-based and shape the rules of conduct for AI business actors. There are three groups of AI business actors: developers, providers and users.

The Guidelines are based on three assumptions:

(1) The use of AI is seen as a solution for some social issues, such as the reduction in the labor force caused by declining birthrates and an aging population.
(2) AI technology is developing rapidly, and its application in society is characterized by speed and complexity. In contrast, the formulation and enforcement of relevant laws tend to be slow.
(3) Rule-based A regulatory approach that clearly stipulates detailed obligations may hinder technological innovation.

The Guidelines adopted a soft laws regulatory approach without any legally binding force that would encourage interested parties to make voluntary efforts to reduce societal risks posed by AI and promote innovations and application of AI.[7] Due to the varying risks associated with AI applications in different fields, the Guidelines provide the guides to the measures to be taken by companies based on a risk-based approach.[8]

---

[4] AI Guidelines for Business Version 1.0, 04. January 2025, https://www.meti.go.jp/shingikai/mono_info_service/ai_shakai_jisso/pdf/20240419_9.pdf.
[5] AI Guidelines for Business Version 1.0, aaO (Fn. 4), S. 11.
[6] AI Guidelines for Business Version 1.0, aaO (Fn. 4), S. 13-21.
[7] AI Guidelines for Business Version 1.0, aaO (Fn. 4), S. 2.
[8] AI Guidelines for Business Version 1.0, aaO (Fn. 4), S. 3.

b) AI Developer

To deal with the matters related to AI developers, Guidelines are divided into three parts and provide measures for each part. These three parts are respectively "during data preprocessing and training", "when developing AI", "after developing AI". During data preprocessing and training developers should properly collect training data and take measures to control the quality of the data. Regarding the fact that biases cannot be completely eliminated from the process of training data, developers should make sure that AI models are trained with properly data sets.[9] When developing AI developers should consider the lives, bodies, properties and minds of humans and the environment. They should consider the possibility that bias can be included by each technical element and make sure AI models are trained with sufficiently representative data sets. They should take security measures appropriately based on the characteristics of the adopted technologies. They should preserve work records for follow-up verification and take measures to maintain and improve the AI quality.[10] After developing AI considerations in each step of development should be identified in order to address the risks of new attack methods to AI systems. Developers should provide the information to relevant stakeholders, such as possibility of changes in output or programs due to learning by AI systems, the expected scope of use set by AI developers in which AI can be safely used. They should prepare documents with development-related information in order to improve traceability and transparency.[11]

c) AI Provider

AI providers are responsible for adding value to the AI system. It is important for AI providers to conduct the appropriate change management, configuration management and service maintenance works. Guidelines on this point are divided into two parts, namely "When implementing an AI system" and "after an AI system or service starts to be provided". When implementing an AI system, the providers should take measures to prevent AI from causing any harm to lives, bodies, property, and mental health of stakeholders and the environment. They should use AI within the expected scope of use set by AI developers and examine how AI usage differ from those that AI developers expect. They should consider bias in the configurations and data of AI systems and services. They should take measures to protect privacy. They should prepare documentation explaining the system architecture and data processing of the AI that influences the decision-making.[12] After an AI system or service starts to be provided, AI providers should periodically verify whether the AI system or service is used for proper purposes.

[9] AI Guidelines for Business Version 1.0, aaO (Fn. 4), S. 27.
[10] AI Guidelines for Business Version 1.0, aaO (Fn. 4), S. 27-28.
[11] AI Guidelines for Business Version 1.0, aaO (Fn. 4), S. 28-29.
[12] AI Guidelines for Business Version 1.0, aaO (Fn. 4), S. 32-33.

They should properly collect necessary information concerning AI systems and services. Regarding the new attack methods, they should identify trends in the latest risks and matters. They should provide the information to the relevant stakeholders.[13]

d) AI Business User

For AI business users, it is important to use AI systems or services properly within the scope of the use set by the AI providers. In addition, human intervention allows human dignity and autonomy to be preserved, helping to prevent unexpected incidents. In line with these principles, the Guidelines provide recommendations on aspects such as "proper use", "consideration for bias", "implementation for security measures".[14]

2. In Discussion

Currently there are the draft "Discussion Points" and the draft "Basic Act on the Advancement of Responsible AI" in discussion.

a) Draft Discussion Points

On May 22, 2024, the AI Strategy Council, which is a government advisory body established to consider the benefits and risks of AI, submitted draft "Discussion Points" concerning the advisability and potential scope of any future regulation. They identify the following risks as ones that Japan should prioritize: safety, privacy and fairness, national security and crime, property protection, and intellectual property. The draft "Discussion Points" suggest that there may be hard law regulations for high-risk AI, while continuing to use soft law where relevant.[15]

b) Draft Basic Act on the Advancement of Responsible AI[16]

aa) Introduction

In January 2023, the Liberal Democratic Party of Japan launched a project team called "AIPT" on the evolution and implementation of AI. The AIPT submitted the draft "Basic Act on the Advancement of Responsible AI" (hereinafter referred to as the "Draft AI Act"). The Draft AI Act is divided into five parts: promote the responsible

[13] AI Guidelines for Business Version 1.0, aaO (Fn. 3), S. 33-34.

[14] AI Guidelines for Business Version 1.0, aaO (Fn. 3), S. 35-36.

[15] AI Watch: Global regulatory tracker – Japan, 15. January 2025, https://www.whitecase.com/insight-our-thinking/ai-watch-global-regulatory-tracker-japan.

[16] Overview of the Basic Law for the Promotion of Responsible AI, 04. January 2025, https://note.com/api/v2/attachments/download/006badee3e4d847b3a0c92358b2de63a.

use of AI, designation of advanced AI foundation model developers, obligation to develop systems by advanced AI foundation model developers, reporting obligation and supervision and penalty, etc. Part two, in particular, concerns the designation of AI foundation model developers of a certain size/objective as an "advanced AI foundation model developer" (hereinafter referred to as the "model developer"). The following issues are discussed: (1) justification and necessity to regulate developers of foundation models as a target of the regulation; (2) how to evaluate/classify based on "size" and "purpose"; (3) should the designation be made unilaterally or notifications be required beforehand? In the case of designating unilaterally, whether or not the State should be authorized to conduct investigations for the purpose of designation; (4) whether or not to impose penalties against business entities who do not report? (5) geographical scope of the regulation. If a notification obligation is imposed, the target business entity shall submit a notification.

Following the designation in Part two, Part three assigns seven obligations to model developers: (1) conduct internal and external safety verification, such as red team testing, for AI in particularly high-risk areas; (2) share risk information among companies and governments; (3) invest in cybersecurity to protect unreleased model weights; (4) incentivize detection and reporting of vulnerabilities by third parties; (5) adopt a mechanism to inform users when generative AI is used for a particular content; (6) publicly report AI capabilities, limitations, etc.; (7) prioritize research on social risks brought about by AI. However, the standards for these seven obligations are expected to be determined by the private sector, including business associations, to keep up with the speed of technological advancement. This regulatory system, where the government defines what to do and the private sector determines how to do it, is generally known as a "co-regulation" model.[17]

Part four concerns reporting obligation and supervision. Model developers are required to regularly report their status of compliance regarding the requirements set forth in part three to the national government or to third parties. The state shall monitor and review model developers based on the above status report. The state may seek the opinions of relevant parties in the private sector. It shall publish the findings of assessments and, in certain cases, request model developers to implement remedies. The state may request reports and conduct on-the-spot inspections in the event that any model developers fail to comply with obligations, or cause an incident.

[17] Policy Research Institute, Publishing of a working draft of a "Basic Law for Promoting Responsible AI" submitted to a project team of the Liberal Democratic Party of Japan, p.3, 04. January 2025, https://www.aplawjapan.com/application/files/3417/0919/1282/Newsletter_Policy_Research_Institute_004.pdf.

bb) Criticism

Regarding the Draft AI Act, the Business Software Alliance (hereinafter referred to as the "BSA") made its recommendation online public.[18] Regarding the definition and regulatory framework of AI, BSA recommends adopting the OECD's definition of AI to align with international standards. Additionally, a risk-based approach should be implemented to better regulate scenarios where the use of AI may cause significant harm.[19] Legislation on AI should aim to avoid conflicts between regulations, especially since existing laws already cover certain aspects of AI. Therefore, the Draft AI Act should focus on addressing gaps in current regulations. When adopting a risk-based approach, the requirements in Draft AI Act should be limited to high-risk use cases rather than high-risk use areas. For example, low-risk AI applications such as background blurring on video calls, autocorrect, email spam filters, web search engines, and TV show recommendations should not be subject to the Draft AI Act.[20]

Regarding the obligations imposed on model developers in the Draft AI Act, BSA believes the drafters cannot effectively achieve the legislative goal. Instead, the diversity and complexity of the AI value chain should be considered, with different obligations assigned to different roles within the AI value chain. For example, AI system developers can access information about the type of data used to train the AI system, its known limitations, and its intended use cases. In contrast, AI deployers can access information about the specific ways they use the system that affect consumers.[21]

Regarding the provisions in Part three, BSA first recommends avoiding prescriptive transparency and reporting obligations, as existing technical standards for AI testing are still in their infancy. Excessive reporting requirements could lead to information overload and force companies to disclose confidential information.[22] For implementation of the so-called co-regulation BSA recommend considering the following aspects: (1) Verify whether government involvement is necessary to provide explanations to the public, while respecting the autonomy of each company. (2) The supervising authority must provide sufficient explanation of the relevance between the requested information and the purpose of the legislation, and demonstrate that it aligns with the fundamental principles of the legislation.(3) Model developers should have the right to

[18] The Software Alliance on the Draft Basic Law for the Promotion of Responsible AI, 04. January 2025, https://www.bsa.org/files/policy-filings/en04112024bsacmtsrespai.pdf.

[19] The Software Alliance on the Draft Basic Law for the Promotion of Responsible AI, aaO (Fn. 18), p.2-3.

[20] The Software Alliance on the Draft Basic Law for the Promotion of Responsible AI, aaO (Fn. 18), p.3-4.

[21] The Software Alliance on the Draft Basic Law for the Promotion of Responsible AI, aaO (Fn. 18), p.4.

[22] The Software Alliance on the Draft Basic Law for the Promotion of Responsible AI, aaO (Fn. 18), p.4-5.

withhold information that may contain trade secrets, and the supervising authority should treat all information provided by the developers as confidential.[23] Considering trade secrets, information that could jeopardize information or network security, and other proprietary information, BSA recommends that companies establish internal testing mechanisms rather than relying on external testing by third parties.[24]

3. Conclusion

Japan has gradually introduced guidelines for AI governance, which were consolidated into the "AI Guidelines for Business Version 1.0" in 2024. These guidelines are not legally binding; rather, they aim to encourage all stakeholders to voluntarily adhere to the recommendations, thereby mitigating societal risks and promoting AI development collaboratively. However, the draft "Discussion Points" submitted in 2024 suggest the possibility of hard law regulations for high-risk AI while continuing to rely on soft law where appropriate. Additionally, the Draft AI Act, submitted by the Liberal Democratic Party, assigns seven obligations to model developers. This indicates that Japan is increasingly recognizing that non-binding guidelines alone may be insufficient to prevent the improper development of AI.

IV. Singapore

1. Current

Singapore also currently has no law specifically directed to regulating AI. The government published the "Model AI Governance Framework" (hereinafter referred to as the "AI Framework") and the "AI Verify" to guide AI development.

a) Model AI Governance Framework

The AI Framework is based on tow principles : (1) Organizations should ensure that the decision-making process is explainable, transparent and fair. (2) AI solution should be human-centric.[25] It intends to assist organizations to achieve two objectivities: (1) Build stakeholder confidence in AI by responsibly using AI within the organization to manage various risks during AI deployment. (2) Demonstrate reasonable efforts in data

[23] The Software Alliance on the Draft Basic Law for the Promotion of Responsible AI, aaO (Fn. 18), p.5.

[24] The Software Alliance on the Draft Basic Law for the Promotion of Responsible AI, aaO (Fn. 18), p.5.

[25] Model AI Governance Framework, S. 15, 10. January 2025, https://www.pdpc.gov.sg/-/media/Files/PDPC/PDF-Files/Resource-for-Organisation/AI/SGModelAIGovFramework2.pdf.

management and protection by aligning internal policies, structures, and processes with accountability-based best practices. Its content is divided into four sections: (1) internal governance structures and measures; (2) determining the level of human involvement in AI-augmented decision-making; (3) operations management; (4) stakeholder interaction and communication. Each section highlights key factors that organizations should consider and provides case examples for detailed explanation.

b) AI Verify

"AI Verify" is an AI governance testing framework and toolkit designed to help organizations validate the performance of their AI systems against AI ethics principles through standardized tests. "AI Verify" was developed by the Infocomm Media Development Authority of Singapore (IMDA) in consultation with private sector organizations. IMDA has also set up the AI Verify Foundation (AIVF), a not-for-profit foundation to gather expertise from private sector organizations and the global open-source community to develop AI testing frameworks, standards, and best practices.

2. In Discussion

There is a draft "Model AI Governance Framework for Generative AI" (hereinafter referred to as the "Draft AI Framework") in discussion, which developed by the AI Verify Foundation (AIVF) and Infocomm Media Development Authority (IMDA). The Draft AI Framework expands on the existing Model AI Governance Framework and provides detailed guidance to private-sector organizations to address key ethical and governance issues when deploying AI solutions.

3. Conclusion

Singapore has consistently relied on non-legally binding frameworks to guide businesses in building trustworthy AI. However, whether this approach is sufficient to address potential AI-related issues remains a point of concern.

V. Taiwan

1. Current

The Taiwan government published the "Unmanned Vehicles Technology Innovative Experimentation Act" on Dec. 19. 2018 to regulate unmanned vehicles experimentation. The Ministry of Science and Technology also announced the "AI research and

development guidelines" in 2019. The Financial Supervisory Commission announced the "Main Principles and Related Promotion Policies for the Use of Artificial Intelligence (AI) in the Financial Industry" on Oct. 17. 2023.

2. In Discussion

Currently the draft "Artificial Intelligence Basic Act"[26], which was proposed by National Science and Technology Council, is in discussion. Its purpose is to present the fundamental values of AI development, governance principles, and policy directions in the form of a basic law, aiming to promote AI innovation while balancing human rights and risk management. Article 3 outlines seven principles that should be followed in AI development:

(1) Sustainable Development and Well-being: AI development should consider social equity and environmental sustainability. Adequate education and training should be provided to reduce potential digital divides and help citizens adapt to changes brought by AI.
(2) Human Autonomy: AI should support human autonomy, respect fundamental human rights and cultural values, and allow human oversight. It should be human-centered while upholding the rule of law and democratic values.
(3) Privacy Protection and Data Governance: Personal data privacy should be properly protected, minimizing the risk of data breaches by applying the principle of data minimization. At the same time, the openness and reuse of non-sensitive data should be encouraged.
(4) Cybersecurity and Safety: Security protection measures should be established throughout AI research, development, and application to prevent threats and attacks, ensuring the robustness and safety of AI systems.
(5) Transparency and Explainability: AI-generated outputs should include appropriate disclosures or labels to facilitate risk assessment, clarify their impact on relevant rights and interests, and enhance AI trustworthiness.
(6) Fairness and Non-discrimination: AI development and application should strive to minimize algorithmic bias and discrimination risks, ensuring that no specific group is subjected to unfair treatment.
(7) Accountability: AI stakeholders should assume appropriate responsibilities, including internal governance responsibilities and external social responsibilities.

Articles 4 to 17 further set out various policy objectives based on the aforementioned fundamental principles and the government's priorities, including innovation

[26] The draft "Artificial Intelligence Basic Act", 15. January 2025, https://join.gov.tw/policies/detail/4c714d85-ab9f-4b17-8335-f13b31148dc4.

and cooperation, talent cultivation, protection of rights, risk management, utilization of data, adaption of laws/regulations and so on.

3. Conclusion

Since 2018, Taiwanese government ministries have gradually established AI regulations according to their respective authorities and needs. However, only the "Unmanned Vehicles Technology Innovative Experimentation Act" holds legal effect, while the other two are merely guidelines without legal binding force. This suggests that the government has adopted a relatively laissez-faire approach to AI development, allowing the industry to grow with minimal intervention. Nevertheless, whether such an attitude is sustainable for the long-term development of AI remains questionable. Therefore, the draft "Artificial Intelligence Basic Act" recently proposed by the National Science and Technology Council aims to incorporate the fundamental values, governance principles, and policy directions of AI into a legal framework. This effort is commendable. However, as the Basic Act primarily sets out the fundamental principles for AI regulation, further discussion is needed on how to implement its content—specifically, which aspects should be legally regulated and which should be presented as guidelines.

VI. Discussion

1. Regulatory approach

The regulatory approaches adopted by various countries differ, likely due to factors such as governance culture, population size, and each nation's stance on AI development. Unlike the other three countries, China first issued two legally binding regulations before introducing an AI governance framework. This reflects an approach that prioritizes regulating AI applications through law, followed by non-binding governance guidelines to steer industry practices and achieve agile governance. Japan, on the other hand, has taken the opposite approach. It initially issued three sets of guidelines and, in 2024, replaced them with the Guidelines for Business Version 1.0, also aiming for agile governance. However, Japan's ongoing discussions around the draft "Discussion Points" and the draft AI Act reveal a growing awareness of the limitations of guiding AI technology development solely through non-binding principles, acknowledging the need for legal intervention. Singapore has consistently relied on guidelines to direct AI technology development, with the goal of achieving agile governance. In Taiwan, apart from the "Unmanned Vehicles Technology Innovative Experimentation Act", the "AI Research and Development Guidelines and the Main Principles and Related Promotion Policies for the Use of Artificial Intelligence (AI) in the Financial Industry"

are not legally binding. The draft "Artificial Intelligence Basic Act" currently under discussion seeks to incorporate the fundamental values, governance principles, and policy directions for AI development into a legal framework. However, this draft does not extend to concrete regulations for AI technology development.

This highlights that, while the AI governance models of the aforementioned countries differ, they all share the common goal of achieving agile governance — emphasizing that AI regulations should not stifle AI development. This stands in contrast to the European Union's AI Act, which focuses on ensuring AI development does not endanger fundamental human rights.[27] As a result, these countries have largely adopted a guideline-based approach to steer industry players in complying with AI development policies, rather than enforcing legally binding regulations like the EU's AI Act. This difference in regulatory attitude not only affects the methods of regulation but also shapes the content of these regulations. The EU's AI Act categorizes AI into four risk levels: unacceptable risk, high risk, limited risk, and minimal risk. AI systems deemed to pose an unacceptable risk are strictly prohibited, while those categorized as high-risk or limited-risk must meet varying requirements. In contrast, Japan's Guidelines do not ban any type of AI. Instead, they offer recommendations tailored to different entities — developers, providers, and users. Since these Guidelines are not legally binding, they can be easily adjusted to keep pace with rapidly evolving AI technologies. However, the issues with this approach have become evident through Japan's Draft Discussion Points and Draft AI Act. Because the Guidelines lack legal enforceability, their effectiveness relies entirely on voluntary compliance from industry players. When conflicts of interest arise, it raises concerns about whether companies will continue to follow the Guidelines. Therefore, for the long-term development of AI, implementing legally binding regulations may become a necessary step. However, the critical challenge lies in ensuring that such legal frameworks align with the fast-evolving nature of AI technology, preventing excessive restrictions that could hinder AI innovation.

2. Purpose and scope of regulation

Building on the previous discussion, adopting a legally binding AI regulatory framework requires careful consideration of two key issues:

(1) Scope of Regulation: Should the law encompass all AI-related issues, or should it focus solely on specific topics? Given the vast range of AI applications, attempting to regulate every aspect of AI is impractical. Such an approach risks either over-regulation or under-regulation and may fail to address the core technological developments and issues accurately, ultimately falling short of its intended goals. Therefore, clearly defining the legislative

[27] Artificial Intelligence Act, Preamble (3).

purpose and appropriately delineating the regulatory scope under that purpose is crucial to ensuring the law's effectiveness.

(2) Relationship with Existing Laws: Another important consideration is how the new AI law will interact with existing legal frameworks. Since AI spans various fields, it may overlap with current laws, such as personal data protection laws, labor laws, civil law provisions on personality rights, and copyright laws. Lawmakers must decide whether AI should be governed by standalone regulations or integrated into existing legal systems. This is essential to prevent redundant or conflicting legal provisions, which could create confusion about which laws apply.

Addressing these two issues will be fundamental in crafting AI regulations that are both effective and adaptable to technological advancements.

3. Risk assessment

In terms of regulation, countries generally agree on the need to consider AI-related risks, making the determination of risk levels a crucial issue. For example, Japan's Draft AI Act requires AI used in high-risk areas to undergo internal and external safety verification by model developers. However, the BSA argues that the risk does not stem from AI technology itself but rather from how AI is utilized. Taking the healthcare sector as an example, using AI for administrative tasks by healthcare providers is not considered high-risk. In contrast, employing AI to assess health insurance reimbursement or the legality of specialized treatments would be classified as high-risk.[28] The EU's AI Act aligns more closely with BSA's perspective. According to Article 6, AI systems listed in Annex III are generally classified as high-risk. However, if the AI's use does not pose risks to health, safety, or fundamental rights, it is not considered high-risk — meaning the risk level depends on how AI is applied. Nevertheless, some scholars have pointed out a potential oversight in this approach. While AI-powered medical devices are classified as high-risk, AI-integrated consumer products such as smartwatches or fitness trackers are not. Yet, considering the collection and reuse of data, these data sets could be employed to evaluate personal mental health, determine health insurance premiums, or even assess job candidates — risks that may be significantly underestimated under the current framework.[29]

[28] The Software Alliance on the Draft Basic Law for the Promotion of Responsible AI, aaO (Fn. 5), p.3-4.

[29] Mühlhoff/Ruschemeier, Regulating AI with Purpose Limitation for Models, 15. January 2025, https://aire.lexxion.eu/article/AIRE/2024/1/5; Ruschemeier/Bareis, Searching for harmonised rules: Understanding the paradigms, provisions and pressing issues in the final EU AI Act, 15. January 2025, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4876206, p. 27.

4. Misuse of AI Technology

The risks posed by AI technology are not limited to its intended uses; they can also arise from the misuse or abuse of AI. Here are two examples. The first example is deep fake technology. This technology has been infamously used to create fake pornographic videos featuring the faces of celebrities. Fraud syndicates have also exploited deep fakes to produce fake videos aimed at deceiving people and defrauding them of their property. Additionally, AI technology has been misused to generate fake news, spread hate speech, and engage in other harmful activities. Another example involves AI systems used to monitor and evaluate production assembly lines to enhance efficiency. Although the data collected may be anonymized, there is a risk that individual data could be reverse-engineered. Similarly, data gathered from consumer products like smartwatches or fitness trackers can be compiled into databases and repurposed for evaluating personal mental health, determining health insurance eligibility, or even assessing job applicants. Such uses go beyond the original intent of AI technology, potentially causing various levels of harm.

The two types of AI misuse described above highlight different forms of abuse, each causing distinct kinds of harm. Therefore, it is crucial to develop tailored strategies to mitigate the risks associated with each type. For the first type — where individuals use AI technology to create fake information for purposes such as fraud or damaging someone's reputation — this issue is not unique to AI misuse but is exacerbated by it. Many countries currently require AI-generated content to be clearly labeled, allowing recipients to recognize that the information was produced by AI. However, it remains questionable whether simply knowing the content is AI-generated enables recipients to identify it as false information or discern the intent behind it. Additionally, these labels can be removed, further complicating the issue. To address this, potential strategies could include prohibiting the removal of AI-generated content labels, reassessing the adequacy of existing regulations — such as platform liability — and exploring new rules to counteract this form of AI misuse. The second type involves the misuse of AI beyond its intended purpose, such as repurposing AI technology or the data it collects. The key countermeasure here is to prevent individuals or entities from using AI systems or their collected data for purposes beyond their original intent. Establishing strict guidelines for data usage, enhancing data protection measures, and imposing accountability mechanisms could be effective ways to mitigate these risks.

VII. Conclusion

This article explores the courses of AI governance in China, Japan, Singapore, and Taiwan, highlighting their distinct approaches compared to Europe. These countries tend

to focus more on promoting AI technology development, often guiding industry practices through non-legally binding guidelines or frameworks, rather than establishing legally binding regulations, except in a few specific areas. However, considering the potential risks associated with AI technology applications and the dangers of misuse, it seems insufficient to rely solely on non-legally binding rules for effective prevention. Therefore, how to regulate AI technology development through legally binding laws and regulations should be a key point of focus.

List of Abbreviations

ADBU	*Assam Don Bosco University*
AEUV	*Vertrag über die Arbeitsweise der Europäischen Union*
AfP	*Zeitschrift für das gesamte Medienrecht*
AGB	*Allgemeine Geschäftsbedingungen*
AI Act	*Artificial Intelligence*
BDSG	*Bundesdatenschutzgesetz*
BeckOK	*Beck'scher Online Kommentar*
BGH	*Bundesgerichtshof*
BGHZ	*Entscheidungen des Bundesgerichtshofes in Zivilsachen*
BVerfG	*Bundesverfassungsgericht*
BVerfGE	*Entscheidung des Bundesverfassungsgerichts*
Cf.	*Compare*
CFR	*Code of Federal Regulation*
CiTiP	*Centre of IT & IP Law*
CJEU	*Court of Justice of the European Union*
CR	*Computer und Recht*
CRi	*Computer Law Review International*
DAS	*Digital Services Act*
DMA	*Digital Markets Act*
DSGVO	*Datenschutz-Grundverordnung*
E.g.	*For example*
EC	*European Commission*
ECJ	*European Court of Justice*
ECLI	*Europäischer Rechtsprechungs-Identifikator*
ECtHR	*European Court of Human Rights*
EJRR	*European Journal of Risk Regulation*
Et at.	*And others*
Et seq.	*And the following*
EU	*European Union*
EuCML	*Journal of European Consumer and Market Law*
Eur. Rev. Priv. Law	*European Review of Private Law*

EUV	*Vertrag über die Europäische Union*
FAZ	*Frankfurter Allgemeine Zeitung*
FCC	*Federal Constitutional Court*
GDPR	*General Data Protection Regulation*
GPR	*Zeitschrift für das Privatrecht der Europäischen Union*
GRCh	*Grundrechtecharta*
GRUR	*Gewerblicher Rechtsschutz und Urheberrecht*
JIPITEC	*Journal of Intellectual Property, Information Technology and Electronic Commerce Law*
JZ	*JuristenZeitung*
KG	*Kammergericht*
MMR	*Zeitschrift für IT-Recht und Recht der Digitalisierung*
MStV	*Medienstaatsvertrag*
MüKo	*Münchener Kommentar*
NetzDG	*Netzwerkdurchsetzungsgesetz*
NJW	*Neue Juristische Wochenschrift*
NVwZ	*Neue Zeitschrift für Verwaltungsrecht*
OJ	*Online Journal*
OLG	*Oberlandesgericht*
OUP	*Oxford University Press*
OVG	*Oberverwaltungsgericht*
RW	*Rechtswissenschaften*
SZ	*Süddeutsche Zeitung*
U.S.	*United States*
UrhDaG	*Urheberrechts-Diensteanbieter-Gesetz*
UWG	*Gesetz gegen den unlauteren Wettbewerb*
ZEuP	*Zeitschrift für Europäisches Privatrecht*
ZfDR	*Zeitschrift für Digitalisierung und Recht*
ZfPW	*Zeitschrift für die gesamte Privatrechtswissenschaft*
ZRP	*Zeitschrift für Rechtspolitik*
ZUM	*Zeitschrift für Urheber- und Medienrecht*
ZUM-RD	*Zeitschrift für Urheber- und Medienrecht – Rechtsprechungsdienst*
ZusProt	*Additional Protocol*

Index of Authors

Joanna Bryson
Professor of Ethics and Technology at the Hertie School, Berlin.

Patricia García Majado
Assistant Professor and Doctor of Constitutional Law at University of Oviedo.

Margaret Hu
Professor of Law, William & Mary Law School in Virginia.

Tobias Mahler
Professor and deputy director of the Norwegian Research Center for Computers and Law at University of Oslo. Co-founder of the Legal Innovation Lab Oslo (LILO).

Irina Orssich
Head of Sector for Artificial Intelligence Policy, European Commission (Directorate General for Communications Networks, Content and Technology).

Lea Ossmann-Magiera / Lisa Markschies
Associate Researcher, Research Group "Norm Setting and Decision Processes" at Humbolt University Berlin / Weizenbaum Institute.

Tristan Radtke
Postdoctoral researcher, TUM Munich.

David Restrepo Amariles
Associate Professor of Artificial Intelligence and Law, HEC Paris.

Antje von Ungern-Sternberg
Professor of German and Foreign Public Law, State Church Law, and International Law at the University of Trier / Executive Director at the Institute for Digital Law Trier.

I-Ping Wang
Professor of Law at National Taipei University (Department of Law).

SCHRIFTEN DES IRDT

TRIER STUDIES ON DIGITAL LAW

Raue / von Ungern-Sternberg / Kumkar / Rüfner (ed.)

Artificial Intelligence and Fundamental Rights

As AI technologies rapidly reshape societies, they pose both tremendous opportunities and serious risks—especially to fundamental rights. In response, the European Union has enacted the groundbreaking AI Act: a legal framework designed to foster innovation while safeguarding democratic values, human autonomy, and the rule of law. But has this delicate balance been struck successfully?

This volume, emerging from the 2024 annual conference of the Digital Law Institute Trier, critically examines the EU's rights-driven approach to AI regulation. From the perspectives of leading scholars and practitioners—including those directly involved in drafting and implementing the Act—it offers deep insights into the legal, ethical, and technical foundations of the AI Act and its global significance.

With contributions on prohibited AI practices, the risk-based regulatory model, and key obligations like data governance and human oversight, the book explores how AI can be regulated to protect fundamental rights without stifling innovation. The book concludes with a comparative view on AI regulation from the United States and Asia.