A brief commentary by Mireille Hildebrandt*

1 The Architecture of the AIA

My first impression is one of awe.

This piece of legislation ties together a series of relevant concerns about the **impact** of AI systems on (1) **human health and safety** and (2) **fundamental rights**. It does so without reinventing the wheel, taking into account the existing (partly upgraded) legal framework regarding potentially dangerous products (machinery, toys, medical devices, aircraft etc.). This has resulted in a multi-dimensional legal framework for those who develop and provide (import or distribute) these systems, and for those who use them (not being end-users).

1.1 Overview

The architecture of the AIA aims to be as simple as possible, but not simpler:

- It deploys a broad definition of **AI systems** to offer broad protection
- It distinguishes between **high risk** and other AI systems

* Research Professor of 'Interfacing Law and Technology', Faculty of Law and Criminology, Vrije Universiteit Brussel. Full Professor of 'Smart environments, Data Protection and the Rule of Law', Science Faculty, Radboud University, Nijmegen. See https://www.cohubicol.com/about/research-team/#mireille-hildebrandt.

- On top of that it defines four prohibited **AI practices**
- Some specified systems must abide by dedicated transparency rules, even if they are not high risk and not part of a prohibited practice
- The AI Act is not applicable to the military
- No new individual rights are attributed
- Obligations are imposed mainly on providers
- High risk systems are defined as such for threats to safety, health or fundamental rights

The **focus is on high risk systems** and on the requirements they must meet to become available on the EU market (and/or to be put into service and/or to be used). The aim of these requirements (chapter 2 in title III) seems to be to ensure **resilience, robustness, reliability and responsibility**. That means that

- part of the requirements see to it that the **claimed functionality** of these systems is verified, validated and tested before becoming available, while
- other requirements see to it that providers **anticipate and either prevent or mitigate potential impact to health, safety and/or fundamental rights** ensuing from both:
    - use for the **intended purpose (claimed functionality)** and
    - **other use cases (reasonably foreseeable misuse).**

1.2 Terminology

Before diving into some of the issues, some terminological points require attention. The Regulation targets **the provision, putting into service or use of AI systems** (not of 'AI' per se, thus avoiding metaphysical speculation or confusing a research domain with devices or infrastructure).

It **prohibits** a set of four **practices** where AI systems are put on the market, put into service or used, and it **imposes a set of stringent requirements** on providers, importers, distributors and users of what are qualified as **high risk AI systems**. Finally, it imposes **transparency obligations** on four specified AI systems (which, may however, also, be qualified as high risk systems and/or be part of a prohibited practice).

AI systems:
The Regulation generally speaks of **AI systems**, which are defined in art 3(1) jo Annex I in reference to **software (whether or not integrated or connected with hardware)** that generates **output**, based on **human defined objectives**, developed with **specified data-driven and/or code-driven techniques**. The definition clearly intends to have a broad scope. Further narrowing down is done when addressing high risk systems (which form a subset of AI systems).

AI practices:
Four types of AI practices are prohibited in art. 5. The term AI ***practice*** is not defined but seems to refer to the 'the placing on the market, putting into service or use of an AI system' that affects natural persons (art. 5 sub a,b,c) or to the 'use' of specific AI systems (art. 5 sub d).

- **Under a and b** the prohibition concerns (a) manipulation or (b) exploitation of vulnerabilities (b) that results in **physical or psychological harm to a natural person**
- **Under (c)** the prohibition concerns social credit scoring by or on behalf of governments that results in **detrimental or unfavourable treatment of a natural person** (with some additional conditions).
- **Under (d)** the prohibition concerns **the use of a specific technology by law enforcement** with some exceptions. No individualised negative results are required for the prohibition to apply.

High risk AI systems
This refers to AI systems that are:
- a product, or a safety component of a product, covered by legislation in Annex II (this mainly concerns health and safety threats)
- a system referred to in Annnex III (this mainly concerns fundamental rights threats)

Note that the distinction between prohibited practices, high risk systems and 'certain systems' with extra transparency obligations **does not refer to mutually exclusive systems**. A system that is not high risk may nevertheless be part of a prohibited practice, and a system with extra transparency obligations may also be part of a prohibited practice or qualify as a high risk system if e.g. used for recruitment. This is potentially confusing.

2 Issues

2.1 Prohibited practices

Three of the prohibited practices are based on a combination of a certain **intent** (manipulation, exploitation of vulnerable persons, social credit scoring by government) and a certain **result** individualised harm/detrimental or unfavorable treatment (see above).

A. I think that requiring identifiable individualised harm/detrimental or unfavourable treatment is highly problematic. First, because the negative impacts of manipulation, exploitation or social credit scoring are not limited to the level of individual harm; it will often play out at the level of democratic processes (e.g. disrupting the public sphere) and diminish public goods such as freedom of expression, human autonomy and fair treatment despite the fact that individual harm cannot be identified (e.g. chilling effect of certain types of surveillance). Second, tort liability law has demonstrated that it is next to impossible to substantiate and prove such individual harm even when it is probable (such proof will be a requirement for a fine ex art. 71.3(a) of the Act and for private law liability). Though the latter can be 'resolved' by imputing strict liability the former (damage to democracy and public goods) cannot be resolved by way of private law remedies.

B. **The fourth prohibited practice** concerns the use of a **specific technology** (remote real time biometric identification systems) if used **by law enforcement** in **publicly accessible space** (with the specified exceptions). The exceptions may seem reasonable, considering their narrowly defined scope, except for
- the third (art. 5.1(d) under iii), that allows such technologies to be used in the context of an arrest- and surrender order it concerns offences to which maximum punishments of 3 years

or more can be imposed. This is not a proper threshold for 'serious crime', giving too much leeway.

C. Meanwhile it is unclear why 'real time' remote biometric identification systems used by private entities in publicly accessible spaces have not been prohibited (with strict exceptions), considering the far-reaching consequences of misidentification in the case of individual persons (note that high overall accuracy can still result in low precision or recall for targeted persons[1]). See the joint EDPB and EDPS Opinion on the AI Act of June 2021.[2] As the EDPB and the EDPS clarify the use of these technologies may easily result in 'the end of anonymity in these spaces'.

D. It is also unclear why the prohibition of social credit scoring is limited to governments, knowing that the eco-system of data brokers, advertising intermediaries, big tech platforms and the use of recommender systems and other types of targeting has created an impenetrable web of nudge-software that has disrupted democratic processes as well as economic markets. Though much of this is built on pseudo-science and may be called out by the requirements for high risk systems, it is unclear whether they will fall within the scope of the currently defined high risk systems and we clearly don't want to depend on self-assessment by the various actors in this obscure and highly complex ecosystem.

    - In line with that, it seems that emotion recognition AI systems, biometric categorisation AI systems and AI systems that aim for neuro-influencing fall in the same category as social credit scoring, whenever they are used to 'reward' or 'punish' behaviour based on unverified, unvalidated and untested systems that are developed and used behind walls of trade secrets and IP rights. Here again, we cannot assume that self-assessment will do the trick. We need serious countervailing powers to ensure that systems behave in ways that favour both democracy and human dignity. See again the joint EDPS and EDPB Opinion and let's note that the current draft does not even qualify these systems as high risk.

2.2 High risk systems

The choice of systems in Annex III seems **arbitrary**. Point 1 refers to technical systems, whereas points 2-8 refer to contexts. Emotion recognition, biometric categorisation and neuro-influencing should be included under point 1, if not prohibited. Several contexts should be added to points 2-8. **Health-related AI systems** that do not fall within the scope of Annex II should be qualified as high risk under Annex III (think health applications, social networks and data brokers working with health-

[1] A fast, clear and correct explanation of why and how high accuracy can nevertheless coincide with low precision can be found here: Christian Yates, 'Coronavirus: Surprisingly Big Problems Caused by Small Errors in Testing', *The Conversation* (blog), http://theconversation.com/coronavirus-surprisingly-big-problems-caused-by-small-errors-in-testing-136700.

[2] 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) of 18 June 2021: https://edpb.europa.eu/our-work-tools/our-documents/edpbedps-joint-opinion/edpb-edps-joint-opinion-52021-proposal_en.

related AI systems), the same goes for AI systems deployed in the context of **insurance** (life, health, real estate etc.), in the context of energy usage (energy usage data can be used to infer profession, religion, or to predict defaulting on payment; this is not about safety components in critical infrastructure as under point 2), and **housing** (think of discrimination based on ethnicity or gender).

Furthermore, under point 8, which concerns 'Administration of justice and democratic processes', the current proposal only lists AI systems used in the context of the judiciary. Considering the **major impact of the use of these systems** (notably for legal search, dispute resolution, automation of decisions made by public administration, and legal advice and representation by attorneys) **on the administration of justice, the nature of legal protection and the system of checks and balances of the rule of law**,[3] all AI systems whose intended purpose is **to research, apply or decide positive law** should be qualified as high risk. This will ensure that such systems are scientifically validated, verified and tested before being integrated in legal practice in the broad sense of that term.

Art. 10 Data and Data Governance
- how will the use of synthetic data fit the requirements?
- 'free of errors and complete' sounds over the top
- 10(4) is crucial, especially taking into account that the world is in flux, the presumption of art. 42 is therefor outrageous, going against the grain of all that is required in art. 10
- 10(5) seems to provide an additional exception to the prohibition of processing of art. 9 GDPR data, though recital (41) seems to deny this
- paragraph 6 is incomprehensible, though admittedly the Regulation pays too little attention to other stages of machine learning (notably the construction of hypothesis space, choice of language, type of ML etc.), it is not clear whether this is what paragraph 6 refers to

Art. 14 Human oversight
- excellent article but paragraph 4(a) seems to require the impossible 'fully understand the capacities and limitations of the high-risk AI system'. What if even the developers cannot assert this (in the case of deep learning systems)? And what about the fact that 'errors, faults or inconsistencies […] may occur within the system or the environment in which the system operates, in particular due to their interaction with natural persons or other systems' (art. 15.3). Although art. 15 rightly requires resilience and reliability, this does not necessarily mean that those tasked with human oversight will fully understand both the capacities and the limitations. I would propose to rephrase as 'has **relevant understanding** of the capacities and limitations of the high-risk AI system'.

Art. 15 Accuracy, robustness, and cybersecurity
- excellent requirements for high risks systems.
- the performance metric for these systems **should not be accuracy only**. On the contrary, the metrics should include precision and recall, which are far more relevant for affected natural persons, see footnote 1.

Art. 43 Conformity assessment
- high risk AI systems of Annex III should all come under the obligation to involve independent

[3] See on the dangers of integration of such software e.g. Masha Medvedeva, Martijn Wieling and Michel Vols, 'The Danger of Reverse-Engineering of Automated Judicial Decision-Making Systems' [2020] arXiv:2012.10301 [cs] <http://arxiv.org/abs/2012.10301> accessed 5 May 2021.

notified bodies with clear expertise in the domain of fundamental rights impact. Though I understand the hesitation to impose this obligation at this point, it will be crucial to set the tone, preventing providers and users of high risk AI systems from ignoring the risks for fundamental rights (noting they may simply have no clue and certainly no incentive).

I could imagine **a timeline** here (e.g. stipulating that independent notified bodies must be involved within 3 years from the coming into force of the Regulation, meaning they have 3 years to develop the auditing skills to assess impacts on fundamental rights).

2.3 Transparency obligations for certain AI systems

- The obligation to inform natural persons that they are interacting with an AI system is crucial, including the fact that this obligation applies at the level of the design of the system.
- Emotion recognition systems and biometric categorisation systems should preferably, as indicated by the EDPS and EDPB in their joint Opinion, be prohibited, with narrow exceptions for e.g. health. The same goes for systems that engage in neuro-influencing. These systems are based on highly problematic evidence while nevertheless generating potentially highly detrimental output (in the form of decisions or behaviour), resulting in chilling effects, exclusion and disrespect for individuals of groups that do not fit the categorisations that are taken for granted.
- The current categorisation of emotion recognition and biometric categorisation as not necessarily high risk, whereas depending on use or context they may nevertheless qualify as high risk or even be prohibited, is **confusing and unnecessarily complicated**. Those developing these systems should at least be forced (by law) to live up to the highest scientific standards, considering what is at stake in terms of output. **If not prohibited,** they should be added under point 1 in Annex III.
- The Regulation does not pay dedicated attention to **neuro-influencing**. This should be outlawed as a prohibited AI practice (with strict exceptions in the case of mental or physical health, e.g. partial recovery of paraplegic patients[4]).

2.4 Harmonisation

As Veale and Zuiderveen Borgesius argue,[5] the **maximum harmonisation** that is claimed for the Regulation, would mean that the categorisation of systems as high risk will determine positive law throughout all the member states. This is particularly concerning where systems that are not

[4] David A Moses and others, 'Neuroprosthesis for Decoding Speech in a Paralyzed Person with Anarthria' (2021) 385 New England Journal of Medicine 217; Ana RC Donati and others, 'Long-Term Training with a Brain-Machine Interface-Based Gait Protocol Induces Partial Neurological Recovery in Paraplegic Patients' (2016) 6 Scientific Reports 30383.

[5] Michael Veale and Frederik Zuiderveen Borgesius, 'Demystifying the Draft EU Artificial Intelligence Act' <https://osf.io/preprints/socarxiv/38p5f/> accessed 17 July 2021.

qualified as high risk and do not count as part of a prohibited AI practice are not regulated – apart from the dedicated transparency obligations with regard to 'certain systems' as discussed above.

This means that much will depend on **the right categorisation of AI systems as high risk** and **a better articulation of prohibited AU practices** (see above). Unless the Act gets this right, it may actually result in less rather than more protection, as it would basically **legitimate** the development, the making available on the market and the use of all systems that are not qualified as high risk and not part of a prohibited AI practice.

3 Enforcement, remedies, individual rights, oversight mechanisms

It seems that the AIA is mainly or solely designed as an administrative law, focused on oversight bodies and administrative fines, refraining from settling private law liability issues and from attributing rights to natural persons. The private law liability regime will follow in the first quarter of 2022.

I think it a small set of crucial rights should be attributed to natural persons, while also including some collective rights:

A. **The right not to be subject to prohibited AI practices**
B. **The right to object to decisions made by high-risk AI systems**
C. **The right to file an injunction in a court of law, and to mandate that right to an NGO in case one is subjected to prohibited AI practices or to decisions made by high-risk AI systems**
D. **The right of dedicated NGOs to file an injunction in their own name with respect to the rights under A and B**