ALLAI.

#General Approach

09.01.2023

AIA policy analysis

Council General Approach

This policy paper analyses the recently adopted Council General Approach for the European Commission proposal for a Regulation for Artificial Intelligence (AIA).

Author:
Catelijne Muller, LL.M

'In the spirit of compromise' and 'with room for improvement'

During the December meeting of the Telecom Council, the Member States of the EU adopted their 'general approach' regarding the proposal for the Regulation on Artificial Intelligence (AIA). This means that we are roughly half way through the legislative process. The European Parliament will likely adopt its position in March of this year, after which the negotiations between the European Commission, the Council and the EP (Trilogue) will start. What stood out at the Council meeting though, was the fact that a notable number of Member States seemed to accept the approach merely 'in the spirit of compromise' signalling 'room for improvement' on several points.

After an initial review of the general approach, we agree that there is indeed room for improvement. Some of our initial concerns have been addressed, but new concerns have arisen, as the text both strengthens and weakens the AIA. It for example proposes a stronger ban on social scoring, better protection against AI-driven exploitation of vulnerabilities, a fundamental rights impact assessment for high-risk AI and a complaints mechanism. It includes our suggestion to add the notion of *foreseeable use* to the notion of 'intended purpose' for general purpose AI systems. But it also proposes blanket exclusions for AI in the military, defence and national security, an extra layer to filter out high-risk AI, and it deletes crucial AI-systems from the high-risk AI list. Below we provide our initial analysis on the main elements of the Council's 'general approach'.

Proposal for a new AI-definition

In reaction to many discussions around the definition of AI, it is often argued that AI-approaches and -methods that do not pose any serious risk, should not be covered by the AI-definition. It would supposedly make the AIA over-inclusive.

This argument fails for several reasons. First of all, it is not the technique itself that determines the scope of the AIA, but rather the domain in which, or the purpose for which, the technique is used. This is duly reflected in the design of the AIA. To determine the actual effect of the AIA on any given AI system, the definition must be read in conjunction with the rest of the AIA. If an AI practice or system is prohibited or categorised as high-risk (ANNEX II or III) or medium-risk (art. 52), that system cannot be used or only under strict conditions. If not, it can in principle be used without restrictions. Hence, the AIA already makes sure that the least risky AI-systems are not faced with restrictions.

The Council proposes a more general definition of AI in art. 3 and suggests that the Commission further specifies the AI-techniques that are covered by the definition in a separate process, outside the ordinary legislative procedure.[1] We strongly recommend to keep the definition as general as possible, because of the well thought out design of the AIA as described above. Moreover, the separate process of 'implementing acts' for determination of AI-techniques creates too much legal uncertainty.

[1] If this route is chosen, the choice between the type of instrument (delegated act or implementing act) is important. It can be said that the Council usually prefers implementing acts because its procedure allows Member States to directly influence the decision making process, whereas the Commission and the European Parliament usually prefer delegated acts. The latter give the Commission greater independence and the Parliament the right to objection.

Stronger and weaker bans

The proposed amendments to the prohibited AI practices both strengthen some and weaken others.

// Exploitation of vulnerabilities

ALLAI welcomes the addition of 'social and economic status' as vulnerabilities that cannot be exploited by AI. We suggest to also add '(mental) health' and 'susceptibility to addictive behaviour' as vulnerabilities.[2]

// Social scoring

We also welcome the addition of 'private actors' to the prohibition of social scoring. We have been calling for this because many instances of social scoring not only by public but also by private actors already exist, that can negatively impact fundamental rights. While these do not consist of 'generalised' schemes of citizen scoring, examples of which we have seen in China, they do consist of the, often indiscriminate, scoring of an individual's social behaviour or characteristics for numerous purposes, such as social benefits fraud detection, loan or mortgage eligibility etc. For the same reason, we welcome the deletion of the word 'trustworthiness' from the prohibitions, resulting in more legal clarity and less room for circumvention of the prohibition.

On the other hand, our concerns as regards the requirement that the score leads to 'detrimental or unfavourable treatment' remain. This condition places the burden of proof entirely on the individual being scored, who might not even be aware of the scoring taking place.

// Biometrics

Our concerns regarding the limited reach of the AI Act where it comes to the many forms of biometric recognition also remain. Still, only biometric identification that is used by (or for) law enforcement purposes will be prohibited, and only if the identification happens in 'real time' (instead of at a later stage), remotely (instead of up close) and in physical public spaces (instead of for example online or at home). The exceptions to this already narrow prohibition, which could make the prohibition virtually ineffective, also remain untouched.

More importantly however, the reach of the prohibition remains limited to biometric *identification*. As we have argued before, this could leave various very intrusive forms of biometric *recognition* that do not necessarily involve identification, largely un-, or at least under-regulated. Biometric categorisation for example, aimed classifying a person as member of a certain group (e.g. 'Caucasian woman between the age of 40 and 50 with blond hair and blue eyes') even moved down the risk-pyramid to the level of medium risk. Biometric assessment, aimed at inferring a person's behaviour or characteristics (in the AIA referred to as emotion recognition), being perhaps the most criticised form of biometric recognition, bordering on eugenics (which is in fact prohibited in the ECFR, at least in the medical field), also remains medium risk. Both these types of biometric recognition do not

[2] A recent news item on Dutch television covered online gambling sites that use algorithms to supposedly identifying gamers' sensitivity to gambling addiction in order to exploit that vulnerability and have them spend more time gambling (https://eenvandaag.avrotros.nl/item/zorgplicht-online-gokbedrijven-laat-nog-te-wensen-over-mensen-kunnen-nog-altijd-te-lang-doorspelen/ https://eenvandaag.avrotros.nl/item/zorgplicht-online-gokbedrijven-laat-nog-te-wensen-over-mensen-kunnen-nog-altijd-te-lang-doorspelen/)

necessarily involve identification, and will thus remain largely unregulated.

We suggest legislators to consider 'flipping' the approach to biometric recognition, and impose a blanket ban on biometric recognition (except for one-to-one biometric identification) while at the same time adding specific and conditional exceptions for example in healthcare.[3]

Blanket exclusions

The Council suggest to expressly exclude a number of domains from the reach (scope) of the AIA:

- Military, defence & national security
- AI-systems used/developed for scientific research and development
- Any research and development *regarding* AI-systems
- Collaboration with third-country public authorities and international organisations in law-enforcement and the judiciary

// Military, defence and national security

The EU has no regulatory competence in matters of the military, defence and national security. We argue however, that the AIA does in fact *not* regulate any such matter. It regulates a tool, that can be used in many domains, including the military, defence and national security. Regulating tools is exactly what the EU *can* do, not in the least in the spirit of harmonisation. More importantly however, it is especially in the military, defence and national security that the use of AI needs the strongest guardrails.[4] Also no clear and accepted definition for national security exists, which in theory broadens the exclusion even further.[5]

A blanket exclusion for the military, defence and national security, would, in its current wording, allow both prohibited AI practices as well as the uncontrolled use of high-risk AI in these domains. We fear that this could open the door to indiscriminate AI-driven manipulation, biometric recognition and social scoring, and unchecked use of AI in high risk domains under the premise of (e.g.) ensuring national security.

// Research and development

As regards AI systems used in/developed for scientific research and development, we suggest to draw inspiration from the REACH regulation, which holds specific arrangements for when hazardous materials are used for scientific research and development.[6] For AI-systems used in or developed for scientific research, such arrangements could for example include compliance with fewer requirements, prior notification/authorisation by national competent authorities, the use of an "AI R&D exception notice/label", etc.

It is unclear what exactly is meant by 'research and developments *regarding* AI-systems', and what kind of R&D would not be covered by the AIA as a result of this exclusion. Also, there is no mention of 'scientific' in this paragraph, which could indicate that it is meant to exclude private R&D (outside of science). A clarification of this exclusion is necessary to understand its meaning.

// Collaboration with third country public authorities and international organisations

---

[3] Muller C. *et al.*: AIA in-depth #2 | Prohibited AI Practices (2022)

[4] The CJEU has established multiple times that "the powers of public authorities face an insurmountable barrier, namely the fundamental rights of individuals".

[5] Muller C. *et al.*: AIA in-depth | Objective, Scope, Definition (2022)

[6] Guidance in a Nutshell - SR&D and PPORD: https://echa.europa.eu/documents/10162/2324906/nutshell_srd_p-pord_en.pdf/14675e6c-b2cf-4049-81ad-3d1bc41ace6d"

When collaborating with the EU or EU Member States on law enforcement and the judiciary, third country public authorities and international organisations are exempt from the scope of the AIA. We worry that this could lead to circumvention of the AIA in certain cases. A third-country law enforcement agency could for example execute prohibited AI practices in situations that affect EU citizens or EU society. We strongly recommend to reconsider this exclusion.

General Purpose AI

First and foremost, we welcome the fact that the Council proposes to include general purpose AI ("GPAI") in the AIA. We also welcome that the Council has taken up our suggestion to include the notion of *foreseeable use* to the notion of 'intended purpose' in art. 4b paragraph 6: "- *any reference to the intended purpose shall be understood as referring to possible use of the general purpose AI systems as high risk AI systems or as components of AI high risk systems in the meaning of Article 6*".

We however think that proposed process of implementing acts, where the Commission determines the actual requirements that GPAI providers should comply with and how they should collaborate with downstream users, outside of the ordinary legislative procedure, does not appreciate the urgent need to regulate GPAI.

The recent release of ChatGPT (as well as earlier releases (and some withdrawals) of generative models such as Dall-E, Stable Diffusion, Lensa, Bloom, Galactica, LaMDA etc.), has put the issue of GPAI in a more critical light. A large community of experts has been testing ChatGPT for various tasks, such as information retrieval, poem writing and legal drafting (contracts, legal briefs, court documents, judgements and even laws). Others have put it to more critical tests in attempts to e.g. discover biases embedded in the model, find the boundaries of what topics it will address, or prove its lack of understanding of human language. Many of them however also acknowledge that ChatGPT, despite its obvious flaws and risks, shows impressive behaviour, with some even signalling a 'paradigm shift' that could fundamentally change how we communicate, teach, learn, create and work.

Because these GPAI systems are meant to be deployed at scale, amidst an overall tendency towards ever more larger and general models, they can become singular points of failure that radiate flaws (e.g., security risks, bias, inequities, fake content, incorrect content, IP violations, etc.) to countless downstream AI applications. This makes that they should be held to at least the same, but perhaps even higher standards than other AI systems.

We have analysed the requirements for high-risk AI of articles 8-15 AIA in light of GPAI systems in an earlier paper, and have come to the conclusion that almost none of the requirements are impossible to meet by GPAI providers[7]. In fact, some requirements can only be met by the GPAI provider and not by a downstream user. We also found that some particular issues that these systems present, such as environmental impact, the potential to produce inaccurate information and fake news at scale, the potential impact on education, skills and language development, IP infringement, and so on, could warrant to put additional guardrails in place. We suggest to add an additional layer on top of the existing requirements for high-risk AI for GPAI systems. Ideally, the

[7] Muller et al. AIA topics | General Purpose AI (2022)

elements of this additional layer would be part of the AIA, so that the separate process of implementing acts that the Council now proposes would not be necessary. If this is however not feasible, we recommend to use the process of delegated acts rather than implementing acts, as the former gives the Commission more independence and puts the EP on equal footing with the Member States.

Fundamental rights impact assessment

We welcome that the Council proposes to implement a fundamental rights impact assessment as a requirement for high-risk AI in art. 9. Moreover, the Council proposes to consider 'the breach of obligations under Union law intended to protect fundamental rights and serious damage to property or the environment' as a 'serious incident' that must be reported once an AI-system is deployed. This is also a welcome addition.

Extra horizontal layer for high-risk AI waters down the AIA

For the applicability of the requirements for high-risk AI, the Council proposes to apply an additional horizontal 'layer'. This layer aims to filter out AI systems that are 'merely accessory to the action' and thus 'not likely to cause serious fundamental rights violations or other significant risks'. As a result, such systems would not be considered high-risk, despite their listing as high-risk on ANNEX III.

What stands out at first sight, is the Council's assumption that the mere *accessoriness* of the AI-system makes it *unlikely to cause serious fundamental rights violations or other signification risks*. We challenge this assumption by giving one very well-known example that contradicts it: the Dutch childcare benefit scandal.

The Dutch childcare benefit scandal is seen as the largest 'AI-scandal' in Europe. It evolved due to a complex interplay of multiple factors, one of which being a flawed AI-system for fraud prediction. Apart from that, there was heavy political pressure to 'detect fraud before it happened', leading to internal pressure to fast-track the deployment of a system unfit for purpose. There also was supposed 'institutional racism' within the tax authority. And there a general lack of understanding of workings of the AI-system and a tendency to take the system's predictions at face value, which led to Kafka-like situations where families were constantly targeted and mistrusted by the tax authority. Because of this combination of factors, it has indeed been argued that the AI-system was merely an 'accessory' to the scandal. We disagree with that argument on principle. But more importantly, the AI fraud prediction system, accessory or not, contributed to 14.000 children being taken from their homes, one suicide, severe (mental) health problems with many affected people, bankruptcies, job losses and ultimately the resignation of the Dutch Government. The proposed 'horizontal layer' could lead to a situation where an AI-system such as the one used in the Dutch childcare benefit scandal, would not be classified as high-risk and would not have to meet any of the requirements that protect exactly those fundamental rights that were affected in the scandal.

The Council also proposes (again) a process of implementing acts through which the Commission would 'specify the circumstances where the output of AI systems referred to in ANNEX III would be 'purely accessory' to the action.

We argue that these circumstances have already been specified in art. 7 of the AIA. We strongly believe that ANNEX III in its current form is the result of a thorough exercise, where the Commission already evaluated potential 'accessoriness' of AI, by applying the filters of art. 7. This article determines the circumstances to be taken into account when adding AI-systems to ANNEX III (and we suppose that the Commission has also taken these elements into account when developing the current version of ANNEX III). The Council basically proposes that the Commission repeats this entire exercise, but then through a much more opaque process outside of the ordinary legislative procedure. This proposal would merely kick the can down the road, opening a window for extensive lobbying and continued legal uncertainty.

An extra horizontal layer also leaves too much room for interpretation and legal uncertainty in the application phase of the AIA. We expect extensive legal discussion on whether an AI system was 'accessory' or not, in legal proceedings. We also fear more stifling of innovation, when it remains uncertain whether such innovation might be covered by the AIA or not.

ANNEX III deletions

As mentioned above, the Council proposes to delete biometric categorisation from the high-risk AI list of ANNEX III. We refer to the paragraph on 'Stronger and Weaker Bans' above for our argumentation against this.

It is our strong recommendation *not* to delete crime analytics (paragraph 6 sub (g)) from the ANNEX III (as the Council now suggests). This paragraph involves 'crime analytics of natural persons by searching complex related and unrelated large data sets available in different data sources in order to identify unknown patterns or discover hidden relationships in the data'. While this practice might indeed help law enforcement, it touches on critical fundamental rights such as the right to a reasonable suspicion and the presumption of innocence. The 'unknown patterns' or 'hidden relationships' that AI systems might help to uncover when combining multiple databases, can be arbitrary and are not necessarily indicators of a crime committed by the suspect. Recent revelations have shown that the Dutch (decentralised) Government is using these types of approaches to predict welfare fraud based on arbitrary indicators and patterns such as 'single mom', 'child returning to live with parents after a certain age', etc.

While we agree that AI-systems used in life- and health insurance should be considered high-risk, we would argue that they are already included in paragraph 5 of ANNEX III: 'Access to and enjoyment of essential private services and public services and benefits.' We do however recommend adding health- and wellness apps and wearables (fitness trackers, period trackers, fertility apps etc.) to ANNEX III. Many of these systems might not be considered medical devices as defined in the MDR and would thus not be covered by ANNEX II. As such they would fall outside the scope of Chapter 2 of the AIA, while they do involve the analysis of biometrics and other sensitive data and often render (semi-)medical advice to individuals. Having these systems comply with the requirements for high-risk AI would ensure that they are of high quality and trustworthy.

Requirements for high-risk AI

The Council proposes to clarify a number of the requirements for high risk AI, which

we welcome. We do however want to express our concerns regarding the absence of any specific requirements for models and algorithms, that also comprise AI-systems. The current requirements focus for a large part on data-driven AI systems (machine learning), but lack specific requirements for logic- and knowledge based systems. We are currently preparing a separate expert policy paper together with AI-scientist on this topic, which will include recommended requirements.

Press inquiries:
welkom@allai.nl
www.allai.nl

ALLAI is an independent organisation that advocates responsible development, deployment and use of AI. ALLAI was founded by the three members of the EU High Level Expert Group on AI, Catelijne Muller, LLM., Professor Virginia Dignum and Professor Aimee van Wynsberghe. ALLAI advises (inter)national governments, organisations and businesses on the ethical, legal and societal implications of AI and its co-founders are Rapporteurs, members and/or advisors of various international organisations (GPAI, OECD, CAHAI, High Level Expert Group on AI, EESC, UNICEF). ALLAI executes different programs aimed at incorporating responsible AI in society, from awareness raising and knowledge building and to responsible AI implementation.