ALLAI.

20.04.2020

#3a

AIA in-depth

#3a | High-Risk AI Classification

Articles 6, 7 & ANNEXES II, III

This report is the third in a series of in-depth analyses of the European Commission proposal for a Regulation for Artificial Intelligence (AIA).

Authors:
Catelijne Muller, LL.M
Noah Schöppl
Maria Avramidou
Christofer Talvitie
Mónica Férnandez Peñalver

CONTENTS

EXECUTIVE SUMMARY

In this paper, which is the third in a series, we will dive deeper into the main elements of Chapter 1 of Title III of the AIA: Classification of High-Risk AI (articles 6 and 7 and ANNEXES II and III AIA). We evaluate the methodology used to classify AI-systems and area's as high-risk, and the implications of those classifications.

As with our previous AIA in-depth papers, our evaluation is guided by the main objective of the AIA: the protection of health, safety and fundamental rights from the ill effects of AI. Central question for this objective is whether AI brings sufficient benefits and is indeed ready to largely replace or heavily influence human decision making, even in critical areas such as law enforcement, the judiciary, social benefits, education or the workplace.

At the time of publication of this paper, the Slovenian Presidency of the European Council issued a proposal for a compromise text for articles 1 - 7 of the AIA. This paper will take this compromise text into consideration where relevant.

Main findings and recommendations

1 **Classification criteria and future proofing:**

- The criteria that lead to a high-risk classification are limited. Some criteria are prioritized hence excluding other (relevant) ones. This is contrary to our fundamental rights doctrine
- Adding new AI-systems to the 'high-risk list' is only allowed in pre-determined domains, making the AIA less 'future proof'
- Not just the intended purpose of the AI-system, but also its 'reasonably foreseeable use' should be taken into consideration

2 **Harmonized products with AI (ANNEX II):**

- We see no reason to exclude the harmonized sectors of ANNEX II.B from the scope of the AIA

3 **Stand-alone high-risk AI (ANNEX III):**

- Biometric identification (one-to-many), categorisation and assessment should be moved to art. 5 AIA
- Telecom, internet, financial infrastructure as well as air, rail and water traffic management should be added to para. 2 as critical infrastructures
- AI driven personalised education should be added to para. 3
- Certain AI(-driven) decisions in employment, e.g. on hiring and termination, should be moved to art. 5 AIA
- AI determining or predicting the (un)lawful use of public services (e.g. fraud risk prediction) should be added to para. 5
- Clarify what kind of private services are to be considered 'essential' (e.g. housing, internet, telecom, financial services, insurance) (para. 5)
- Predictive policing, criminal profiling and biometric lie detection in law enforcement, criminal justice and asylum, migration and border control should be moved to art. 5 AIA
- AI to make judicial decisions should be moved to art. 5 AIA and 'the judiciary' should be clarified in para. 8
- AI used for vote counting in elections should be added to art. 5 AIA
- Content moderation in democracy-critical processes should be added to para. 8

1 CLASSIFICATION OF HIGH-RISK AI

According to the AIA, AI-systems that pose a risk of harm to health and safety or fundamental rights of persons, but nevertheless (supposedly) bring benefits to society, are considered high-risk. Those AI systems will be allowed on the Union Internal market, as long as they comply with a set of mandatory requirements and follow conformity assessment procedures before they can be placed on the Union market.

The chosen approach first and foremost poses the risk of normalizing and mainstreaming quite a number of AI practices that are still heavily criticized, often due to their lack of (evident) sufficient social benefit. Secondly, this approach assumes that the risks high-risk AI systems pose, also future ones, can be sufficiently mitigated by meeting the requirements of articles 9 to 15 AIA (which are (paraphrased: risk management system, datasets of high quality and data governance, documentation, transparency, human oversight, accuracy, cybersecurity and robustness).

But perhaps the most pertinent question this approach raises is: are we ready to allow high-risk AI to largely replace or heavily influence human decision making, even in critical processes such as law enforcement or the judiciary or at the workplace? And is AI ready?

1.1 Criteria for high-risk classification

In art. 7 paragraph 1 the AIA lays down the categorisation rules for high-risk AI. This article is written for future high-risk AI systems that should be brought into the scope of the high-risk requirements of the AIA, but it also gives an idea of how the Commission initially determined the domains and AI systems that are currently considered high-risk (and listed in ANNEX II and III). The Commission's aim is to keep the AIA flexible and forward looking, which is commendable given that AI is developing fast and new applications and techniques are constantly emerging. We do, however, see some areas for improvement of this approach.

1.1.1 **Limited to 8 pre-determined domains**

First and foremost, only AI systems or practices that fall within the domains already described in ANNEX III can be added to the list. Moreover, there is no mention of new harmonized rules that could justify AI(-driven) products to be considered high-risk in the future. A truly visionary approach would allow for new AI systems or practices to be considered for the high-risk AI lists, regardless of the domain.

1.1.2 **Limited criteria**

Secondly, the AIA lays down 8 criteria that the Commission should take into account to assess whether a new AI system poses a 'risk of harm to health, safety and fundamental rights' and should be added to the high-risk list. First of all, the Commission can only add AI systems to the high-risk list that pose a *similar or greater* harm to health, safety and fundamental rights than the AI-systems already mentioned in ANNEX III. Further, the AIA sets pre-conditions for what exactly would be considered harm and, as such, prioritizes certain circumstances over others that might nevertheless be relevant. For example, paragraph 2 sub (d) sets 'the ability of a system to affect a plurality of persons' as a condition. This could limit the possibility (and the need) to consider the harm AI can bring to a small number of persons or even to one person as a condition. Also, it is unclear whether the conditions are meant to be limited or merely serve as inspiration, whether they are cumulative or facultative, or whether they should be looked at in connection with each other.

Lastly, this approach undermines legal doctrine where our fundamental rights are to be interpreted and applied considering all relevant circumstances, without predetermined thresholds or boundaries for what can be considered a relevant criterion or determinant for such risk. The pre-conditions of art. 7 paragraph 1 and the criteria of art. 7 paragraph 2 could have an explanatory purpose if described in the recitals of the AIA. Putting them in the body of the AIA in their current wording creates a legal imbalance with our fundamental rights doctrine.

1.1.3 **Intended purpose**

In paragraph 2 sub (a) the AIA describes the *intended purpose of the AI system* as a condition. According to the Commission, the classification of an AI system as high-risk does not only depend on the function performed by the AI system, but also on the specific purpose and modalities for which that system is used. By doing so it limits the possibility to consider the '*reasonably foreseeable use*' of the system as a criterion. This can (and already seems to have) create(d) a potential loophole. The Slovenian Presidency compromise proposal excludes so-called 'general purpose AI systems', the argument being that for providers of these systems, it would be impossible to know all intended purposes their system could be used for.

In our first in-depth paper, AIA in-depth #1 | Objective, Scope, Definition, we already argued that excluding 'general purpose AI' bears too many risks, as these could become single points of failure that resonate throughout downstream uses. We thus suggested adding the notion of 'reasonably foreseeable use' to the notion of 'intended purpose'. Given the potential impact of these 'general purpose AI systems' it is not unreasonable to ask from their providers to try to foresee (imagine) the potential uses of the AI system and address the risks to health, safety and fundamental rights these uses could bring. Several AI experts have already argued that 'general purpose AI systems should not be excluded from the scope of the AIA for these reasons.

RECOMMENDATION

Article 7
Amendments to ANNEX III

1 The Commission is empowered to adopt delegated acts in accordance with Article 73 to update the lists in ANNEX II and A and III (...). ~~where both of the following conditions are fulfilled:~~
(a) DELETE
(b) DELETE

2. When assessing (...), that is equivalent or greater than the risk of harm posed by the high-risk AI systems already referred to in ANNEX III, the Commission shall take into account, including but not limited to, the following criteria:

(a) The intended purpose or reasonably foreseeable use of the AI system;

2 HARMONIZED PRODUCTS WITH AI

The AIA considers AI that is a safety component of a harmonized product listed in ANNEX II part A, or is itself such a product, to be high-risk. The ‘interwoven system’ between the AIA and Union Harmonisation Legislation is a good approach from a lawmaking point of view and avoids doubling of governance and accountability structures and obligations.

Art. 6 AIA specifically refers to AI safety components, or AI that is itself a product, in order to target components of products, which may affect the safety of individuals. The definition of safety components is to be interpreted in a broad manner, so as to include all components the failure or malfunctioning of which endangers the health and safety of persons or property, according to the Commission in its reaction to the EESC opinion INT/940. The Commission clarified that for example AI systems which perform diagnostic or therapeutic functions are high-risk. These systems might not be considered ‘safety components’ in a literal sense or be themselves products (but only part of a broader diagnostic process), they should nevertheless be considered high-risk. To increase clarity, article 6 could be slightly amended so as to include this broad interpretation of ‘safety component’.

RECOMMENDATION

Article 6
Classification rules for high-risk AI systems

1. Irrespective (...)

(a) the AI system is intended to be used as a ~~safety~~ component the failure or malfunctioning of which endangers the health, safety or fundamental rights of persons or the ~~safety~~ of property (...);

(b) the product whose safety component as meant under (a) is the AI system (...).

2.1 ANNEX II part A versus part B

ANNEX 2 holds two parts, A and B. AI-systems that are part of products listed on part A are considered high-risk while those that are part of products listed on part B are excluded from the scope of the AIA entirely. Part B includes domains that deal with 'wings, wheels, water and rails'. According to Recital (29), the Commission deems it more appropriate to amend the already strict EU regulations and directives that underpin these domains rather than have these domains be covered by the AIA.

This reasoning is confusing, because other domains that are also covered strict regulations and directives, such as medical devices, lifts, toys etc., listed on part A, are expressly considered high-risk and not excluded from the scope of the AIA. To avoid interference with existing requirements and procedures in these domains, the AIA simply states that the requirements for high risk AI of the AIA are to be included in those existing procedures. One would assume that the same structure could be followed for AI in relation to 'wings, wheels, water and rails'.[1]

[1] Muller et. al (2022): AIA in-depth #1 | Objective, Scope, Definition

3 STANDALONE HIGH-RISK AI

Annex. III (standalone high-risk AI) lists a number of AI uses in a total of 8 areas that are considered high-risk. According to the Commission, these systems bring enough social benefit to justify their use, provided that a number of requirements are met. In the next paragraphs we analyze the domains and uses to determine and weigh their benefits and risks.

High-Risk AI or Unacceptable AI Practice?

Many of the AI-systems and practices on ANNEX III border on, and sometimes even involve, practices that could be (and perhaps should be) prohibited under art. 5 AIA. These practices particularly involve biometric recognition and social scoring. The use of AI systems that score or people, e.g. on their creditworthiness, fraud risk, job suitability, benefit eligibility, risk of domestic violence, risk of child abuse, risk of poverty, criminality, reliability, future job performance, skills, personality, growth potential, recidivism risk, etc. has grown exponentially over the past couple of years. Depending on the data used to make these predictions, some of these practices could be considered a form of unacceptable social scoring. [2]

[2] In AIA in-depth #2 | Prohibited AI Practices we suggest a clarification of the prohibition of social scoring, so as to draw a clear line between what is to be considered unacceptable social scoring and what can be considered legitimate assessment.

3.1 Biometric identification and categorization

In AIA in-depth #2 | Prohibited AI Practices [3], we assessed that art. 5.1 (d) prohibits only a very narrow set of biometric recognition practices for an even narrower group of actors. As an alternative, instead of banning only a narrow set of biometric identification practices and categorizing all others high and medium-risk, our paper echoes the various calls (e.g. by EDPS, EDPR, EESC, IBM, and a number of MEPs) for a ban on biometric recognition (which includes biometric identification, but also all forms of biometric categorization and assessment both by private organizations and (semi-)public authorities). The call for such a ban also resonates more and more among the wider public.

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

If, in exceptional instances (such as for example those already mentioned in art. 5.1 (d) subparagraphs (i), (ii) and (iii)), biometric identification is considered, efficacy (proportionality and necessity) should be well determined and established and measures should be taken to limit its use as regards time and/or scope. Also, such use should always be subject to Title III of the AIA (high-risk AI systems). Moreover, the risk of misuse where these systems need to be in place, even if they are only 'switched on' only in certain circumstances, should be effectively mitigated. Another situation where some of these AI practices could bring benefits is in controlled environments such as for example hospitals, where the technology could serve a scientific purpose. Here also, we warn that these uses should at a very minimum be evidence based, limited in time, proportionate and necessary, meaning that the exceptional character of these exceptions should be specified, for instance by defining a maximum amount of time or share of space for the use in these exceptional cases. This regulatory mechanism has for instance been used in EU state aid rules to make sure that exceptions are not exploited as loopholes. Biometric verification or authentication (one-to-one matching) remains allowed. Also here, however, it remains crucial to make sure that these uses do not allow for circumvention or create loopholes.

3.2 Management and operation of critical infrastructure

Also here, the text refers to ‘safety components’. As described above, according to the Commission the definition of safety components is conceived in a broad manner, so as to include all components the failure or malfunctioning of which endangers the health and safety of persons or property. This could be included in this paragraph in the same manner.

Missing from this ‘area’ are AI systems used in the management and operation of the telecom, internet and financial infrastructure and the management, generation and supply of electricity and energy (including nuclear power).

RECOMMENDATION

Alternative text for paragraph 2 ANNEX III:

2. Management, operation, generation and supply of critical infrastructure, technology and energy:

   (a) AI systems intended to be used as a component, the failure or malfunctioning of which endangers the health, safety or fundamental rights of persons or the safety of property, in the management, operation, generation and/or supply of the telecom, internet, and financial infrastructure, road, rail, air and water traffic, and the operation, management an/or supply of water, gas, heating, and electricity and energy (including nuclear power).

3.3 Education and vocational training

AI systems to determine access to education and evaluate students, in particular where these applications use biometrics and behavior recognition, pose a number of risks of harm to student health, safety and fundamental rights.

Algorithmic grading used during the pandemic caused serious uproar in the UK as it incorrectly downgraded students having detrimental effects on their life opportunities. Online proctoring tools track multiple things such as eye movement, mouse movement, keyboard strokes, soundscapes, copy and paste behavior, online search behavior, body movement, etc. to supposedly flag ‘suspicious behavior’ and ‘indications of cheating’ during online exams. Scientific evidence of the effectiveness of these techniques is generally absent, while students have experienced the use of these systems as invasive and having a negative effect on their ability to conduct the exams in a dignified manner. These systems mostly involve AI-driven biometric assessment, for which we refer to our second in-depth paper, AIA in-depth #2 | Prohibited AI Practices [4], where we suggest prohibiting these practices. Hence, some AI practices in education and vocational training should be banned, such as AI proctoring, while some should be classified as high-risk, such as algorithmic grading.

Lastly, other practices such as individualized AI teaching systems, which aim to optimize learning processes based on learning data of students, should also be added to the high-risk list in the education section.

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

RECOMM.

Addition to paragraph 3 of ANNEX III:

3. Education and vocational training:

   (c) AI systems intended to be used for the optimization of individual learning processes based on a student's learning data.

3.4 Employment, workers management and access to self-employment

The AIA adds certain AI-systems in the workplace, such as for recruitment, promotion, termination, task allocation, and monitoring and evaluating performance and behavior of workers, to ANNEX III. Many of these AI-practices are also referred to as ‘algorithmic management’ and involve a plethora of AI-systems and tools that (help) monitor, track, assess, evaluate, recruit, hire and dismiss workers. It is important to understand that it has several ‘layers’ to it. It involves tools that track workers’ activities, both on- and offline. Think of eye-tracking, the tracking of mouse movement and clicks, room noise, typing speed, but also the tracking of physical movements, such as walking speed, driving routes, etc. And it involves tools that, based on these and other inputs, make or inform decisions about workers. These decisions can involve rostering, or the allocation of tasks, but also productivity scoring, promotion, demotion, and even termination. During the Corona crisis, we have seen a surge in AI-driven applications aimed at monitoring worker productivity while working from home.[6]

[6] Home Office (Surveillance) – ALLAI

While the high-risk classification sounds logical, but this approach can easily lead to the legitimization, normalization, and mainstreaming of quite a number of work-related AI-applications for which the actual impact on workers, but also on existing labor laws, is still not fully understood. This impact should not be underestimated. After all, the AI-system's activities described concern the fundamentals of employment relationships: monitoring, evaluation, assessment, pay, promotion, and dismissal, to name just a few. Fundamentals that have been the topic of laws, rules and standards for many years. We must ensure that the AIA does not actually weaken the position of the employee or undermine existing regulation. Particularly in labor relations, where there is a power imbalance, these types of systems should not be implemented without proper worker consultation and the involvement of social partners. Hence, it should be seriously considered to move certain algorithmic decisions in the employment sector, such as those involving hiring or termination, to the unacceptable risks of art. 5 AIA.

Many of the AI-systems used in recruiting and ‘algorithmic management’ consist of ever more intrusive forms of worker or applicant surveillance, monitoring and scoring, often by using special categories of personal data (biometrics) or irrelevant or unrelated information. For our in-depth analysis of biometric recognition (in all its forms, including biometric surveillance, categorisation and assessment) and social scoring, please refer to our second in-depth paper, AIA in-depth #2 | Prohibited AI Practices, where we clarify the prohibitions so as to also include these practices.[7]

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

3.5 Access and enjoyment of essential private services and public services and benefits

Between 2013 and 2019 the Dutch tax authorities employed an AI-system to predict and classify the risk of fraud with child benefits. The Dutch tax authorities used a multitude of ‘risk classifiers’, one of them being the non-Dutch nationality. As a consequence authorities wrongly accused an estimated 26,000 parents and caregivers, from mostly low-income families, of making fraudulent benefit claims.

The authorities suspended their benefits and required them to pay back the allowances they had received in their entirety and immediately. In many cases, these claims amounted to tens of thousands of euros, driving families into severe financial hardship. Besides, the people flagged by the system as 'fraudsters' were often subjected to ‘kafka-esk’ investigations, characterized by harsh rules and policies, rigid interpretations of laws, and ruthless benefits recovery policies. The Dutch Privacy Authority considered the use of nationality in the fraud risk classification model discriminatory[8]. The Parliamentary Interrogation Committee considered the consequences an “unprecedented injustice" and the Dutch Government eventually resigned over the scandal.

[8] Belastingdienst/Toeslagen: De verwerking van de nationaliteit van aanvragers van kinderopvangtoeslag

For the access and enjoyment of public services, the paragraph includes AI-systems that assess eligibility to receive services and benefits as well as systems that decide or inform decisions on the revoking or reclaiming of services and benefits. In light of the Dutch childcare benefits scandal, we urge that this description is meant to be interpreted in a broad manner so as to also include AI-systems that perform benefit fraud prediction, risk analysis etc.

We also advise to slightly broaden the scope of this paragraph beyond natural persons, and include the self-employed and (other) micro-enterprises. These groups are often considered one-person businesses and the impact of high-risk AI can be as grave for them as for natural persons.

As regards access to and enjoyment of essential private services, recital (37) indicates that these include access to finance, housing, electricity and telecommunication. They should however at the very least also include insurance, energy and the internet. Moreover, it should be noted that when it comes to private services, only AI-systems aimed at credit(worthiness)-scoring are currently considered high-risk. AI is however being used not only to assess creditworthiness in relation to private services, but also for example to assess the risk of fraud, the amount of premiums, etc, which should be included as high-risk as well.

It should be noted that all these systems border on social scoring. Please refer to our second in-depth paper, AIA in-depth #2 | Prohibited AI Practices [9], where we clarify social scoring so as to draw a clearer line between unacceptable social scoring and legitimate assessment.

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

Alternative text for paragraph 5 ANNEX III:

5. Access to and enjoyment of essential private services and public services and benefits:

(a) AI systems intended to be used by or on behalf of (semi-)public authorities or private parties to evaluate or predict the lawful use by, or the eligibility of, natural persons, including the self employed and micro-enterprises, for public assistance, benefits and services and essential private services including but not limited to housing, electricity, heating/cooling, finance, insurance and internet, as well as to grant reduce, revoke, or reclaim such benefits and services or set payment obligations related to these services;

(b) DELETE

RECOMMENDATION

3.6 Law enforcement

Many of the listed AI uses in the area of law enforcement involve crime prediction or profiling (6. (a) (e), (f), (g)). Others (can) involve forms of biometric categorisation and assessment (6. (b)). Virtually all pose a significant risk of harm to multiple fundamental rights (sometimes even to multiple or all at the same time). Human dignity, privacy, the right to non-discrimination, the presumption of innocence, the right of defence, and the right to a fair trial all become vulnerable with the use of these types of AI systems.[10]

[10] Muller C. (2019): ALLAI advises Council of Europe on AI & Human Rights, Democracy and Rule of Law, for the Council of Europe's CAHAI

For instance, lie detection and emotion detection with biometric recognition (6. (b)) is scientifically flawed and truly invasive. So far, lie detection test results have been inadmissible in European courts. Nevertheless there has been research ongoing on AI-based lie detection tests for border control, funded by the European Commission.

Many AI systems that make risk assessments regarding individual criminality or criminal intentions and the likely occurrence of crime etc. (predictive policing) merely seek correlations between characteristics that a person or an area happens to share with other convicted criminals or areas (such as address, income, nationality, debts, employment, behavior, the behavior of friends and family members and so on). This is incompatible with the presumption of innocence and the right to reasonable suspicion, because the system does not make its prediction based on an actual suspicion of the suspect. Predictive policing has also led to undesirable feedback loops where the same communities are surveilled more often than others, invading the right to privacy and often also the right to non-discrimination. Moreover, it is impossible for legal professionals to understand the reasoning behind the outcomes of AI-driven predictive systems, which affects the right to a fair trial and a reasonable defense. The same goes for AI used in the criminal justice system to predict or profile defendants, determine sentences, etc.

[10] See also: http://partnershiponai.org/wp-content/uploads/2021/07/Why-PATTERN-Should-Not-Be-Used.pdf and https://www.fairtrials.org/app/uploads/2022/03/Ban_Predictive_Policing_Criminal_Justice_Statement.pdf

This begs the question whether the use of a technology or practice in a domain where it (often simultaneously) invades multiple fundamental rights should even be allowed in the first place. We strongly recommend substantially limiting the use of these types of AI by or for the purpose of law enforcement and in criminal justice processes by adding them to the prohibitions.[11] Multiple organizations have called for a ban on predictive policing and profiling in the criminal justice system.

RECOMMENDATION

We recommend to add the following AI practices listed in ANNEX III point 6, which involve predictive policing, criminal profiling and biometric assessment, to the prohibitions of article 5 AIA: (a), (b), (e), (f), (g) and broaden the notion of 'law enforcement' so as to prohibit these practices in the entire 'criminal justice system'.

3.7 Migration, asylum and border control management

According to a recent paper by the EPRS: “The EU’s centralized information systems for borders and security are increasingly incorporating biometric technologies for the purpose of identity verification or identification, automated fingerprint identification. Facial recognition is expected to be used by all systems except one in the near future. A number of EU-funded projects and initiatives have explored and piloted emotion recognition technologies at the EU border.”

As with AI used in law enforcement and criminal justice, AI used in migration, asylum and border control management for making individual (criminal, security or health) risk assessments or performing lie detection, infringe upon the same fundamental rights including the right to asylum. Moreover, border control technologies can be used as a testing ground for developing surveillance capabilities that can later be used on the general population. The fundamental rights of those EU citizens and non-EU citizens passing EU borders must be protected.

Art. 83 of the AIA however initially excludes AI systems which are components of these systems from its scope, if they are put into service before the application of the AIA. This enables EU's centralized information systems for borders and security to implement prohibited and high-risk AI systems in the near future, only having to comply with AIA when the entire system is evaluated.[12]

[12] AIA in-depth #1 | Objective, Scope, Definition

RECOMMENDATION

We recommend to add the following AI practices listed in ANNEX III point 7, which (criminal) profiling and biometric assessment, to the prohibitions of article 5 AIA: (a) and (b).

3.8 Administration of justice and democratic processes

The use of AI in the administration of justice and democratic processes is particularly sensitive and should be approached with more nuance and scrutiny than is done now.

3.8.1 **Administration of justice**

We argue that a distinction should also be made between criminal justice on the one hand and civil and administrative justice on the other. As regards criminal justice, we refer to paragraph 3.6.

There have been a lot of developments and an increase in the use of AI in the legal sector ranging from AI to perform purely administrative and organizational tasks (such as the allocation of cases), and AI that performs contract drafting and legal research, to AI that performs predictive analytics that informs litigators, prosecutors and judges in procedural strategies, judicial decisions and sentencing [13].

[13] Schwermer et al.

According to ANNEX III, AI-systems to assist a judicial authority in researching and interpreting facts and the law and in applying the law to a concrete set of facts are to be considered high-risk. First of all, the limitation to AI-systems used to assist a 'judicial authority' causes uncertainty as to what is meant by such judicial authority. Would AI used by a public prosecutor or district attorney be included? And what about AI used by complaints boards, in arbitration or in online dispute resolution? The AIA should be amended to also include private arbitration and online dispute resolution in the high-risk category.[14]

[14] See: Nigel Stobbs, Dan Hunter and Mirko Bagaric. 2017; Can Sentencing be Enhanced by the Use of Artificial Intelligence? *Criminal Law Journal*, 41, 5, 261–277; https://lexmachina.com/what-we-do/how-it-works/; Ravel Law, http://ravellaw.com/products/#eluid7fcb4be2

Secondly, the text refers to 'assistance' of the judiciary, which leaves fully automated judicial decision making out of scope. The argument here will expectedly be that fully automated decision making is already prohibited under the GDPR. This is however only partially true for two reasons. First the GDPR only prohibits decisions that are based *solely* on automatic processing. This condition has led to different interpretations and uncertainty as to when this is the case. Second, the prohibition holds several exceptions, such as consent, Member State authorization, or the necessity to enter into or perform a contract. These exceptions could also be relied upon in legal processes, in particular when it comes to complaints boards, in arbitration or in online dispute resolution.

The work of virtually all legal professionals (judges, attorneys, solicitors, barristers, company lawyers, legal advisors, arbitrators etc.) is extremely complex, multi-faceted and context dependent. At the same time it is crucial for the upholding of the rule of law. Since we have seen examples where AI systems that assisted the judiciary that turned out to be ineffective, prone to bias and even harmful, these systems need serious scrutiny. Even for simpler AI systems that perform case searches, a recent study has shown that different systems can have very disparate outcomes, showing only 7% of the same outcomes. This also underlines the fact that AI is human made and the biases and assumptions of those who construct these systems determine the results.[15]

[15] Susan Nevelow Mart, 2016. The Algorithm as a Human Artefact: Implications for Legal {Re}Search. 1–50

Recital (40) states that AI systems in the judiciary (and in democratic processes) should be high-risk ‘considering their potentially significant impact on democracy, rule of law, individual freedoms as well as the right to an effective remedy and to a fair trial’ and with the purpose to addressing ‘the risks of potential biases, errors and opacity.’ This indicates that the Commission envisions the broad protection of people’s fundamental rights when they are involved in legal processes.

There is one additional area of concern not covered by this paragraph, which is the practice of translating laws and the interpretation and application thereof into AI systems that then perform the interpretation and application of such laws ('rules as code'). This area goes beyond the scope of the mere administration of justice and can venture into various public and private domains, where the correct and fair interpretation and application of the law is crucial. We argue that these AI processes should also be considered high risk.

Amendment of paragraph 8 ANNEX III:

8. Administration of civil and administrative justice (...):

(a) AI systems systems to be used for or assist in civil or administrative legal disputes and proceedings researching and/or interpreting facts and/or the law and/or in applying the law to a concrete set of facts and/or taking decisions with legal effect.

(b) Processes where legally binding rules are translated into AI systems (rules (as code).

Moreover, we strongly recommend to add AI to make judicial decisions to the prohibited AI practices of art. 5 AIA

RECOMMENDATION

3.8.2 **Democratic processes**

While the ‘header’ of paragraph 8 also mentions ‘democratic processes’ as a high-risk area, it lists no specific AI-systems or uses. This could indicate that the Commission envisions all AI-systems in democratic processes to be considered high-risk, however, this is not entirely clear.

The use of AI has shown to be able to influence democracies in unintended and unexpected ways. The most widely discussed example of this is Cambridge Analytica, which exploited data of 87 million Facebook users to build profiles that could be utilized for political gain. But we need to look beyond Cambridge Analytica to understand the full exposure of our democracies and societies to AI.

The distortion of democracies, public discourse, social cohesion, and public trust by AI cannot be pinpointed to one event, scandal nor even a single phenomenon. AI-driven computational propaganda has distorted elections in Ukraine, Estonia, China, Iran, Mexico, the UK, and the US.. Moreover, Facebook's algorithm has incited violence against protesters in Myanmar, and whistleblower Frances Haugen has said that the company's algorithm is "fanning ethnic violence" in Ethiopia. These are just a few contemporary examples of how AI can impact our democracies and social cohesion, but we cannot even begin to fully grasp how AI could potentially shake democratic societies in the future.

In our second in-depth paper, AIA in-depth #2 | Prohibited AI Practices, we already proposed to amend the first two prohibitions of art. 5 paragraph 1 sub a) and b) to prohibit AI-driven harmful manipulation and exploitation.

Apart from that we also recommend to add AI applications for vote counting in elections, to the prohibited AI practices. Vote counting processes depend on society being able to trust them and even witness and understand the process in detail. A centralized technologically complex system - even if its accuracy can be verified by experts - would risk losing the common trust in and understandability of the electoral process.

We moreover propose to add other AI-systems and practices that can be used in the realm of elections and democratic processes as well as in election news aggregation and provision to the high-risk list of ANNEX III.

Recommendations of AI & democratic processes

We recommend to prohibit the use of AI for vote counting and add it to art. 5.1

We recommend the following amendment to ANNEX III:

8. (...) and democratic processes:

(b) AI-systems intended to be used in elections and democratic processes as well as in election news aggregation

RECOMMENDATION

3.9 Missing: content moderation

Missing from the high-risk list is the area of online content moderation, and in particular AI used for the detection of online fake news, deep fakes, an otherwise harmful content (violence, discrimination, abuse, etc.). While AI could provide a partial solution towards detecting and eliminating harmful online content, it struggles to correctly detect fake news, due to the lack of political, cultural and social knowledge and common sense, and thus poses the risk of censorship.

Moreover, the Cambridge Analytica scandal compounded fears that the algorithms that determine what people see on the platform were amplifying fake news and hate speech, and that Russian hackers had weaponized them to try to sway the election in Trump's favor. Large scale incidents such as the Cambridge Analytica have shown the impact that content moderation could have to democratic rights and values. The use of AI for this purpose should be added to ANNEX III.

ABOUT ALLAI

ALLAI is an independent organisation that aims to foster, promote and achieve the responsible development, deployment and use of AI.

ALLAI's mission is to take a holistic approach to AI, taking into account all impact domains such as economics, ethics, privacy, laws, safety, labour, education, etc. ALLAI aims to involve all stakeholders in its mission: policy-makers, industry, social partners, consumers, NGOs, educational and care instructions, academics from various disciplines.

ALLAI was founded by the three Dutch members of the High Level Expert Group on AI, Catelijne Muller, LLM, Prof. Virginia Dignum and Prof. Aimee van Wynsberghe.

ALLAI refers to Stichting ALLAI Nederland, a foundation under Dutch Law. No entity or person connected to ALLAI, including its Board Members, Advisory Board Members, employees, experts, volunteers and agents, is responsible or liable for any direct or indirect loss or damage suffered by any person or entity relying wholly or partially on this communication.

CONTACT

ALLAI
Prinseneiland 23A
1013 LL Amsterdam
The Netherlands

www.allai.nl
@ALLAI_EU
welkom@allai.nl