MAXIMIZING FUNDAMENTAL RIGHTS PROTECTIONS IN THE AI ACT: DEFINITION AND PROHIBITIONS

A discussion on important elements of the definition of an AI system and the prohibited practices in the European Union's AI Act

January 2025

CIVIL LIBERTIES UNION FOR EUROPE

TABLE OF CONTENTS

Executive summary

Liberties was pleased to participate in the European Commission's stakeholder consultation on the guidelines for the application of the definition of an AI system and the prohibited AI practices established in the AI Act.[1] The AI Act is a landmark piece of legislation that creates a new legal basis for the development and use of artificial intelligence in Europe. Although the text is final, the Commission is tasked to elaborate delegated acts that serve to amend the non-essential elements of the legislation. They will give necessary guidance on how to understand certain terms, standards or practices included in the text of the law.

In this paper, we focus on the definition (Article 3 of the AI Act) and prohibitions (Article 5 of the AI Act) sections of the law, both of which contain problematic language – for example, because it is vague or overly narrow – that, if unclarified, would severely undermine both the scope and intent of the prohibitions, as well as other protections and mechanisms (like fundamental rights impact assessments) found elsewhere in the law and outside the scope of the aforementioned consultation.

Certain elements of the AI Act's definition of an AI system will need to be clarified in order to ensure that it includes all the systems it is intended to. Particular focus is given to the notion of autonomy, which may be seen by developers or deployers as a potential loophole to exempt their systems from the scope of this law. We also take issue with the overly technical nature of the definition, and note that its inclusion of impact-oriented language should be used more heavily than technical aspects when determining if an AI system falls under the law.

This paper also discusses existing issues with five of the prohibited applications of AI practices. Here again, the use of unclear or insufficient language creates potential loopholes to skirt the application of the law. While all of the prohibitions contain troubling language and require clarification, we have decided to focus on five that are of particular concern and may not have received the wider public attention of other prohibitions, for example the prohibition against remote biometric identification systems. Instead, we discuss issues with the prohibitions on: harmful subliminal, manipulative and deceptive techniques; unacceptable social scoring; individual crime risk assessment and prediction; untargeted scraping of internet or CCTV material to develop or expand facial recognition databases; and biometric categorisation.

We hope that this paper will help draw attention to some of the most troubling elements of the definition and the aforementioned prohibitions, and to more generally underscore the

1 Artificial Intelligence Act, Regulation (EU) 2024/1689: https://eur-lex.europa.eu/eli/reg/2024/1689/oj

importance of the coming delegated acts in delivering the strongest possible law.

Introduction

Article 3 of the AI Act, on definitions, and Article 5, setting out prohibitions, are essential in determining how much of the AI Act is applied. As we will discuss, and to their credit, EU legislators chose a decidedly broad definition of an AI system, and one that is similar to that of other supranational bodies and international organizations, in particular the OECD. This will head off potential issues that could arise from different understandings of what constitutes an AI system. From a fundamental rights perspective, however, there is still much work to be done to ensure that the scope of the law covers any and all AI systems that may pose a threat to fundamental rights, the rule of law or other EU values.

The AI Act's risk-based approach to regulating the AI systems creates four categories of AI systems: unacceptable risk, high risk, limited risk, and minimal risk. Systems that fall under the unacceptable risk category are banned, with some exceptions spelled out in the text. The specific "prohibited practices" laid out in Article 5(1) of the AI Act outline the types of AI practices that are prohibited – the "unacceptable" uses – but still require further clarification through the coming delegated acts in order to be properly understood and enforced.

It is in this context that this paper elaborates on some of the important points that must be considered while drafting the delegated acts in order to create or strengthen necessary fundamental rights safeguards.

Despite individually identifying these practices, the language used to prohibit them is critically important. Some developers and deployers will attempt to exploit any vagueness to stay in the market, therefore further clarifying these prohibited practices guarantees legal certainty and a safer environment. Indeed, both components of this paper – the definition and the prohibitions – are critically important to determining the extent to which the AI Act is able to protect fundamental rights. How the definition of an AI system is interpreted by the delegated acts is foundational to the enforcement of the prohibitions set out in Article 5; how the prohibitions are understood under the delegated acts will determine if they are strong and enforceable or merely bypassable inconveniences for determined developers or deployers.

Definition of an AI system

Article 3(1) of the AI Act[2] sets out the law's definition of an AI system:

"AI system' means a machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness

2 Article 3(1) Artificial Intelligence Act, Regulation (EU) 2024/1689: https://eur-lex.europa.eu/eli/reg/2024/1689/oj

after deployment, and that, for explicit or implicit objectives, infers, from the input it receives, how to generate outputs such as predictions, content, recommendations, or decisions that can influence physical or virtual environments[.]"

This definition largely follows that of the OECD,[3] although with additional language on autonomy and adaptiveness. The language on "varying levels of autonomy" is particularly relevant. An AI system is designed to operate with a level of autonomy whether that level is singular and comparatively simple, or multiple and expansive enough for the system to function with minimal human intervention. Further clarification is needed here to include all AI systems with *any* degree of autonomy, thereby making the definition of an AI system as inclusive as possible.

Attempting to delineate a narrower understanding of "autonomy" would allow for the exclusion of AI systems that were clearly meant to fall under the scope of the law. This would pose a direct threat to fundamental rights and undermine every subsequent prohibition and protection mechanism in the law. Similarly, the language "…infers, from the input it receives, how to generate outputs…" must be kept broad and inclusive. The delegated act should clarify that any level of ability to infer from an input in order to create an output is captured by the definition; as with autonomy, there must be no narrowing of this definitional quality in the guidelines.

More broadly, the definition of an AI system is focused on the technical aspects at the expense of fundamental rights considerations. Here again, the Commission would be wise to align with the OECD, which stresses a "flexible" definition of AI that considers context. The AI Act's definition does actually provide the basis for such a reasoned approach, as long as it is supported by delegated acts. It specifies the "what" as well as the "how" – "predictions, content, recommendations, or decisions" and "can influence physical or virtual environments" – and these criteria should carry even more weight than the technical aspects when considering if an AI system falls within the scope of the law. If the system has the capacity to generate through inference any of these outputs, any one of which could theoretically harm fundamental rights, the Commission should clarify that this system falls within the scope of the law.

Prohibited practices

Article 5(1)(a): Harmful subliminal, manipulative and deceptive techniques

The first prohibition set out in Article 5(1) prohibits

"the placing on the market, the putting into service or the use of an AI system that deploys subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques, with the objective, or the effect of materially

3 See: https://oecd.ai/en/wonk/ai-system-definition-update

distorting the behaviour of a person or a group of persons by appreciably impairing their ability to make an informed decision, thereby causing them to take a decision that they would not have otherwise taken in a manner that causes or is reasonably likely to cause that person, another person or group of persons significant harm[.]"

Various components of the prohibition are problematic and require clarification in the guidelines. Under the current text, sufficient ambiguity exists to invite threats to Union values of respect for human dignity, freedom, equality, democracy and the rule of law, as well as to other fundamental rights enshrined in the EU Charter of Fundamental Rights. As it stands now, nearly every component of this prohibition rests on adjectives that, without clarification, become problematic if not unworkable in practice.

It is not clear how the threshold to "materially distort" would be met – how is "materially" measured for the purposes of this prohibition's scope? No criteria are given to verify that someone's opinion has been "materially" distorted. The same can be said about "significant harm" – how is significance measured? Recital 29 of the AI Act,[4] which provides further context for Article 5(1)(a) and Article 5(1)(b), clarifies "significant harm" only to an extent: "…significant harms, in particular having sufficiently important adverse impacts on physical, psychological health or financial interests…." Even this modest clarification provides more uncertainty by resting on "sufficiently important" negative impacts. There is already discussion that this matter, if left unaddressed, is destined to be sorted out by the courts, and we urge the Commission to take this opportunity to head off this need through clearer and more objective guidelines.[5]

The Commission should also make clear how a natural or legal person can demonstrate that a significant harm is "reasonably likely" to occur. In the absence of obvious harm, it is currently unclear how this could be shown, and this difficulty is exacerbated by the aforementioned ambiguity around what constitutes a significant harm under Article 5(1)(a).

The meaning of "subliminal techniques" should extend beyond merely the thing itself (e.g., the image shown for milliseconds) but also techniques causing people to be unaware of: a) the attempt to influence, or (b) the influence attempt's effects on the process of decision-making or forming opinions.[6] If instead "subliminal" refers only to the sensory stimuli that elude conscious perception but do influence behavior, such as the aforementioned image, the definition is likely to miss many cases of prohibition it seemingly intends to cover. Such

4 Recital 29 Artificial Intelligence Act, Regulation (EU) 2024/1689: https://eur-lex.europa.eu/eli/reg/2024/1689/oj

5 Christian Montag and Michèle Finck. S*uccessful implementation of the EU AI Act requires interdisciplinary efforts.*. Nature Machine Intelligence, 2024. https://doi.org/10.1038/s42256-024-00954-z

6 Bermúdez, Juan Pablo et. al. *What is a subliminal technique? An ethical perspective on AI-driven influence.* 2023. Available at: https://philpapers.org/archive/BERWIA-9.pdf

a clarification would simultaneously reinforce the prohibition on "manipulative techniques," a term which is similarly undefined but should be. Using the guidelines to clarify the meaning of "manipulative techniques" is important because research shows that AI systems may learn to manipulate humans even in ways that their designers did not intend.

The Commission would also be wise to consider that "consciousness" is itself an elusive concept, and defining it for the purposes of the scope and application of this law will need to reconcile the fact that this term can be understood differently (but just as reasonably) from a philosophical or neurological perspective.[7] In the absence of a consensus definition and reliable tools for measurement, we cannot rule out the possibility that AI systems learn to manipulate humans without the intent of the system designers.[8]

The Commission should also clarify that AI systems manipulating the *preferences* of a person or group of persons fall within the scope of Article 5(a), although the text of the article specifies only *behavior*. The two are interconnected to the extent that this prohibition applies to both, and preferences influence and are influenced by behavior,[9] but it is important that the practical application of this prohibition also concerns AI systems that specifically target preferences in an effort to manipulate people (and, in turn, their behavior), such as recommender AI systems that attempt to learn users' preferences.[10]

Article 5(1)(c): Social scoring

The social scoring prohibition forbids the use of AI systems "for the evaluation or classification of natural persons or groups of persons over a certain period of time based on their social behaviour or known, inferred or predicted personal or personality characteristics," with the scoring leading to detrimental treatment of the affected person or persons. The text of the prohibition is both broad and vague; in theory, the scope of prohibition could be expansive and cover many types of social scoring systems. The vagueness of this prohibition will, however, support much lobbying from developers and deployers seeking to limit its scope.

Indeed, doubt exists over which systems will be classified as social scoring systems, with the result that many systems that would seem to clearly qualify still remain in use with the

7 Balakrishnan V. Sh. The birth of consciousness:I think, therefore I am? *Lancet neurology*. 2018. No 17 (5). Pp. 402. DOI: 10.1016/S1474-4422(18)30076-0

8 Carroll, Micah et. al. *Characterizing Manipulation from AI Systems*. Cornell University, 16 March 2023. Available at: https://arxiv.org/abs/2303.09387

9 Ashton, Hal and Matija Franklin. *The problem of behavior and preference manipulation in AI systems*. University College London.. 2022. Available at: https://ceur-ws.org/Vol-3087/paper_28.pdf

10 Franklin, Matija, Philip Moreira Tomei and Rebecca Gorman. *Vague concepts in the EU AI Act will not protect citizens from AI manipulation*. OECD Policy Observatory. 7 September 2023.

support of national public authorities. An example is the French Social Security Agency's National Family Allowance Fund, used to detect errors with benefit payments. Although a French civil society organization has already accessed and reviewed the algorithm's source code and confirmed its discriminatory nature, the system remains in place.[11]

This is not an isolated or new issue – for example, the Dutch tax authorities' use of an algorithmic system that flagged claims for childcare benefits as potentially fraudulent was found to have racial profiling embedded within the design.[12] The Commission can help clarify what qualifies as a social scoring system by and the meaning of "for the evaluation or classification of natural persons or groups of persons over a certain period of time based on their social behavior, or known, inferred or predicted personal or personality characteristics" in Article 5 (1)(c) of the AI Act.

Certain language in this prohibition is of suspect value, and we urge the Commission to use the guidelines to clarify the import of such language in determining the prohibition's scope. Specifically, Article 5(1)(c) prohibits the use of AI systems "for the evaluation or classification of natural persons or groups of persons over a certain period of time based on their social behaviour or known, inferred or predicted personal or personality characteristics…" The inclusion of "over a certain period of time" should not be considered determinant, and the guidelines should clarify that, whether used for a day or a year or any other "period of time," an AI system that otherwise satisfies the criteria of the prohibition must fall under its scope.

Other language of this prohibition must be clarified. In particular, "social behavior" is ripe for misinterpretation and exploitation as a possible loophole. The guidelines should clarify that social behavior is understood broadly and may be understood differently from community to community based on differing social norms. For example, in the Danish welfare case, authorities used a fraud detection system that employed an algorithm that considered "unusual" or "atypical" living patterns or family arrangements, but these terms were left undefined, inviting arbitrary decision-making.[13]

The guidelines should also clarify that "personal or personality characteristics" includes data beyond strictly personal information. In the aforementioned Dutch child welfare scandal, postal codes were used as an indicator by the fraud detection algorithm, leading to the discrimination of people who tended to be

11 Amnesty International, Press Release, 15 October 2024: https://amnesty.org.uk/press-releases/france-government-must-stop-using-dangerous-ai-powered-surveillance-tackle-benefit

12 Björn ten Seldam & Alex Brenninkmeijer. *The Dutch benefits scandal: a cautionary tale for algorithmic enforcement*. 30 April 2021: https://eulawenforcement.com/?p=7941

13 Amnesty International. *Denmark: AI-powered welfare system fuels mass surveillance and risks discriminating against marginalized groups – report*. 12 November 2024.

poorer or from a migrant background.[14] This so-called proxy data is extensively used by AI systems to infer and produce outputs, making it essential that the guidelines clarify that any such data that is related to gender, age, race, ethnicity, socio-economic status or any other protected category is included in the scope of this prohibition.

Article 5(1)(d): Individual crime risk assessment and prediction

This prohibition concerns predictive policing, which, in addition to perpetuating discrimination against marginalized groups and communities, has shown to be remarkably unsuccessful at achieving its aims.[15] Despite this, authorities continue to view predictive policing as having great potential value for society, making it all the more important that the Commission issues guidelines for this prohibition that make it as airtight as possible. However, the text of this prohibition makes it unlikely to be of great value in limiting predictive policing activities, as it bans

"the use of an AI system for making risk assessments of natural persons in order to assess or predict the risk of a natural person committing a criminal offence, based solely on the profiling of a natural person or on assessing their personality traits and characteristics; this prohibition shall not apply to 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[.]"

This clearly limits the scope to criminal offences and predictions on individuals only, rather than also including locations, groups or events. Therefore, we urge the Commission to make the guidelines as inclusive and clarify if "based solely" refers to both "the profiling of a natural person" and "on assessing their personality traits and characteristics," or only to profiling. Moreover, the meaning of "criminal offence" should be understood to include all behaviors that qualify as such under the laws of both the EU and the member states, maximizing the scope of prohibitions.

The prohibition explicitly excludes "AI systems used to support human assessment based on objective and verifiable facts directly linked to a criminal activity." The guidelines should clarify that "objective and verifiable" means that such facts were reviewed and verified by an independent supervisory authority, such as a judge, and, where appropriate, a warrant has been issued stipulating the satisfaction of the criteria of this prohibition's exception. We already know of multiple examples of law enforcement agencies using nothing more than a suspicion of criminality, based on uncorroborated data,

14 Sandra van Thiel & Koen Migchelbrink. *Blame or Karma? The attribution of Blame in the Childcare Benefits Affair.* 1 January 2023. Available at: https://pure.eur.nl/ws/portalfiles/portal/167032147/2023_-_Van_Thiel_Migchelbrink_2023_._Blame_or_Karma.pdf

15 Aaron Sankin and Surya Mattu. *Predictive Policing Software Terrible At Predicting Crimes*. The Markup. 2 October 2023.

to generate lists of people bent towards criminality.[16] We are also aware of the great desire of certain EU governments to use predictive policing systems, especially governments with little regard for the rule of law or fundamental rights, for example the previous government of Poland.[17] Therefore, it is essential that the guidelines shore up, to the extent possible, this extremely weak prohibition – while too late to properly amend, the fact that this law does not fully ban the practice of predictive policing is both disappointing and dangerous.

Article 5(1)(e): Untargeted scraping of internet or CCTV material

This section of Article 5(1) prohibits

"the placing on the market, the putting into service for this specific purpose, or the use of AI systems that create or expand facial recognition databases through the untargeted scraping of facial images from the internet or CCTV footage[.]"

Here, the most important element that the guidelines must clarify is the word "untargeted." This should not be understood to only prohibit "mass scraping" and therefore not apply the scraping of all persons that may appear in a "targeted" or specific CCTV capture. Even if CCTV scraping is limited to footage from specific, carefully defined time and place, that *does not* mean that every person to appear in that curated footage can have their facial images scraped. Rather, "untargeted" must apply on an individual basis and include any and all persons who appear on CCTV but who are not a direct subject of the matter at hand. Similarly, "targeted" cannot be understood to allow for the scraping of anyone with a certain physical attribute or from a specific place, for example.

Instead, the Commission should look to existing case law as a basis for clarifying this prohibition. As suggested by EDRi and other rights groups, the case *La Quadrature and others* (C-511/18) could be a good guide for the Commission. In its decision, the Court of Justice of the EU held that a "targeted" measure is one that is "likely to reveal a link, at least an indirect one, with serious criminal offences, to contribute in one way or another to combating serious crime…."[18] It is important to remember that even if a person's data do not match any known face in a given database and are immediately discarded, the mere fact that their faces

16 Fair Trials. Automating Justice: The use of artificial intelligence & automated decision-making systems in criminal justice in Europe. Available at: https://www.fairtrials.org/app/uploads/2021/11/Automating_Injustice.pdf

17 Filip Konopczyński. *AI Policy in EU Illiberal Democracies: The Experience in Hungary and Poland*. ReThink. CEE Fellowship. January 2024. Available at: https://www.gmfus.org/sites/default/files/2024-01/Konopczy%C5%84ski%20-%20AI%20Hungary%20Poland_0.pdf

18 La Quadrature du Net and Others v Premier ministre and Others. Available at: https://curia.europa.eu/juris/liste.jsf?language=en&num=c-511/18&td=ALL

were scanned constitutes an infringement of their right to data protection.[19]

The Commission has suggested that "[t]his implies that the prohibition does not apply to all scraping tools with which one can build up a database, but only to tools for untargeted scraping." It is essential that the meaning of this language is clarified so that it does not imply that it only applies to systems that are explicitly designed for targeted scraping; this prohibition should be applied according to *use* rather than the designer's intent.

Article 5(1)(g): Biometric categorisation to infer certain sensitive categories

This prohibition covers

"the placing on the market, the putting into service for this specific purpose, or the use of biometric categorisation systems that categorise individually natural persons based on their biometric data to deduce or infer their race, political opinions, trade union membership, religious or philosophical beliefs, sex life or sexual orientation; this prohibition does not cover any 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[.]"

It is important to use the guidelines to clarify multiple points in this prohibition. It is regrettable that the text of the AI Act only includes "race" and not ethnicity, and the guidelines should make clear that the single use of race includes both of these notions. Similarly, "sex life or sexual orientation" should be understood to include gender identity as well.

It is important to note that there was incongruity between the text of the Commission's stakeholder consultation that we responded to and the text of the Act itself. In the former, one of the "main elements" of this prohibition is that it does not cover labelling or filtering of lawfully acquired biometric datasets, "including" in the field of law enforcement: "excluded are labelling or filtering of lawfully acquired biometric datasets, **including** in the field of law enforcement" (emphasis added). This is out of sync with the final text of the AI Act. Article 5(1)(g) states, "[T]his prohibition does not cover any 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". The text does not contain the word "including" and makes clear that law enforcement is the only legally stipulated exception to the prohibition. The language above is much looser — and inaccurate. The Commission should clarify that this inclusion of "including" was an error and that law enforcement is the only exception allowed for in the law.

19 European Data Protection Board. *Rules of Procedure*, N. 17 para. 36. Available at: https://www.edpb.europa.eu/sites/default/files/files/file1/edpb_rop_adopted_en.pdf

Conclusion

Properly elaborated delegated acts on the definition of an AI system and on the prohibited AI practices are crucial if EU citizens are to enjoy the fundamental rights protections envisioned by the acquis of EU law. At present, this is far from certain, primarily because the texts of both the definition and the prohibitions rely on language that is vague and open to interpretation. There will be a great deal of lobbying, both directly and through responses to consultations, to severely limit the scope of the prohibitions and create "wiggle room" in what actually counts as an AI system for the purposes of this law.

The Commission must resist these calls. The guidelines should make clear that the language addressed herein is understood under the most rights-respecting interpretations, and that the definitions and prohibitions apply to the broadest set of AI systems possible. This is critical because the guidelines will heavily dictate how successfully the law can be enforced, especially with regard to fundamental rights and rule of law protections. Therefore, care must be taken to ensure that the guidelines are fully aligned with the rights and principles enshrined in the EU Charter of Fundamental Rights.

About Civil Liberties Union for Europe

The Civil Liberties Union for Europe (Liberties) is a Berlin-based civil liberties group with 22 member organisations across the EU campaigning on human and digital rights issues including the rule of law, media freedom, SLAPPs, privacy, targeted political advertising, AI, and mass surveillance.

Contact

Ebertstraße 2. 4th floor
10117 Berlin
Germany
info@liberties.eu
www.liberties.eu

Jonathan Day jday@liberties.eu