AUTORITEIT
PERSOONSGEGEVENS

Call for input

Manipulative, deceptive and exploitative AI systems

Prohibitions in EU Regulation 2024/1689 (AI Act)

Autoriteit Persoonsgegevens (NL) - Department for the Coordination of Algorithmic Oversight (DCA)

September 2024
DCA-2024-01

Summary

The European AI Act (2024/1689) has been in force since 1 August 2024 and regulates the use of AI in the European Union (EU). The AI Act has a risk-based approach. As a result, certain AI systems with an unacceptable risk are prohibited from 2 February 2025.

It is up to market surveillance authorities on the AI Act to provide clarification on how the prohibitions will be enforced. In order to prepare for this in the Netherlands, the Autoriteit Persoonsgegevens (AP) asks interested parties (citizens, governments, businesses and other organisations) and their representatives for information and insights. We can use all input to consider the necessary further clarification of the prohibited AI systems.

In this first call for input, we delineate specific parts of the first two prohibitions: manipulative and deceptive AI systems (prohibition A) and exploitative AI systems (prohibition B). Later we will also ask for input on the other prohibitions. This document outlines specific criteria for these prohibited AI systems, while requesting (additional) input through a set of questions. Contributions can be submitted until 17 November 2024.

The AP makes this call for input based upon its role as coordinating supervisor of algorithms and AI. The AP was given this task in 2023, in addition to its role as Dutch data protection authority. For the purpose of this new task, the Department for the Coordination of Algorithmic Oversight (DCA) was established. This call for input also aligns with the preparatory work being done in support of future supervision of AI systems prohibited under the AI Act. The Dutch government is currently working on the formal designation of national supervisory authorities for the AI Act.

I. Background

1. **The European AI Act (2024/1689) has been in force since 1 August 2024.** This regulation sets out rules for the provision and use of artificial intelligence (AI) in the European Union. The premise of the AI Act is that while there are numerous beneficial applications of AI, the technology also entails risks that have to be managed. The legislation follows a risk-based approach. More restrictive rules will apply for those AI systems that pose a greater risk. Some systems entail such an unacceptable risk that their placing on the market or use is completely prohibited. This is for example the case with AI systems that are used for manipulation or exploitation. The prohibitions are set out in Article 5 of the AI Act.

Prohibited AI applications as from February 2025

2. **The prohibitions in the AI Act will become applicable soon.** As from 2 February 2025, prohibited AI systems may no longer be put on the European market or used. As from 2 August 2025, market surveillance authorities should be designated for prohibited AI systems, and sanctions may be imposed for violations of the prohibitions. Before that time, violation of one of the prohibitions could already lead to civil liability.

Supervision in the Netherlands on compliance with the prohibitions

3. **The Dutch government is currently working on legislation designating which supervisory authority will be responsible for overseeing compliance with the prohibitions.** In the Netherlands, the AP (through its Department for the Coordination of Algorithmic Oversight) and the Dutch Authority for Digital Infrastructure (RDI) provide advice on the supervisory framework for the purpose of the AI Act. They do this in cooperation and coordination with other supervisors. In a second interim advice, published May 2024, the Dutch supervisors proposed to make the AP primarily responsible for the supervision of prohibited AI. In case the recommendations are incorporated, for the prohibitions on manipulative and exploitative AI systems, the AP will closely cooperate with other relevant supervisors, in particular with the Dutch Authority for the Financial Markets(AFM). The supervisory authorities advise the government to designate the AFM as the market authority when prohibited AI systems are used in the AFM's area of supervision. The AP will also cooperate with the Netherlands Authority for Consumers and Markets (ACM) because of the interrelation with the legal provisions on manipulative and misleading practices in consumer law and the Digital Services Act (DSA), among others.

4. **With this call for input, a preparatory foundation is laid for further explanation of the prohibitions in the AI Act.** Because the prohibitions in this call concern AI systems that also fall under other Union laws, this call has been coordinated within the Dutch Cooperation Platform of Digital Supervisory authorities (AI and Algorithm group), and the call has also been discussed in the National Working Group of AI Supervisors, in the spirit of the requirement in article 78(8) of the AI Act to consult national competent authorities responsible for other Union law that cover AI systems.

II. About this call for input

Purpose: Why do we ask for input

5. **It is up to the supervisors of the AI Act to provide clarification on how the prohibitions will be enforced.** In preparation for this, the AP is asking for information and insights from stakeholders (citizens, governments, business and other organisations) and their representatives. All responses can be used for the further explanation of prohibited AI. Within the AP, the Department for the Coordination of Algorithmic Oversight is charged with this task.

6. **This call for input discusses the prohibitions outlined in Article 5, paragraph 1, subparagraphs a and b of the AI Act.** The legislative text and the recitals serve as the foundations for this call for input. Given the scope of the prohibitions in these subparagraphs, this first call for input is limited to these two prohibitions. Please refer to the annex of this document for an overview of the prohibitions in subparagraphs a through g of Article 5, paragraph 1 of the AI Act.

7. **This call for input highlights specific aspects of these two prohibitions.** The focus is on those specific criteria that determine whether or not an AI system is within scope of these prohibitions. Each point is briefly explained based on the legislator's interpretation in the explanatory memorandum to the AI Act. In some cases, we provide an interpretation of our own. We then pose several questions.

Process: this is how you send your input to us

8. **You decide which questions you answer.** You can also provide us with other relevant input in addition to the questions asked. Please send your input by email to dca@autoriteitpersoonsgegevens.nl by 17 November 2024 at the latest. Please mention the topic "Call for input DCA 2024-01, manipulative, deceptive and exploitative AI systems", your name and/or your organisation in your email. If desirable, you can provide us with your contact details so we can reach you when we have further questions. When we have received your input, we will send a confirmation by email.

Follow-up: what will we do with your input?

9. **After the closure of this call for input, the AP will publish a summary and appreciation of the input on manipulative, deceptive and exploitative AI systems**. In this summary, we will refer to the input received in generic terms (e.g., "several sectoral representative organisations have indicated", "a developer of AI systems points out that", "organisations advocating for fundamental rights have noted that"). If preferred and indicated by you, we may explicitly mention your organisation or group. Through our summarised and evaluative response, we can also share the acquired insights with other (European) AI supervisory authorities. For instance, the input received may be utilised in the drafting of the guidelines regarding the prohibitions. At the European level, the AI Office – part of the administrative structure of the Directorate-General for Communication Networks, Content and Technology of the European Commission – can collaborate with the market surveillance authorities to develop such guidelines.

10. **We will only use your input for our task to obtain information and insights about the prohibitions in the AI Act.** We will delete your personal data after publication of our summary and evaluation of the input, unless you have given permission for further use. For more information about how we process personal data, please see: The AP and privacy.

More calls for input are coming up

11. **Following this call, there will be more calls for input on other parts of the AI Act, including the other prohibitions.** In the short term, the AP wants to publish a call for input on prohibition F: AI systems for emotion recognition in the workplace or in education.

III. Definition of the prohibitions on manipulative, deceptive and exploitative AI systems

General scope of prohibited AI systems

12. **The AI Act (and its prohibitions) apply to 'AI systems'.** Thus, in order to determine whether the Regulation applies, an important question is whether the product falls under the definition of an AI system:

> "A machine-based system that is designed to operate with varying levels of autonomy and that may exhibit adaptiveness 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."

13. **The prohibitions are addressed to providers (e.g. developers), deployers, importers, distributors and other operators.** These operators shall not place on the market, put into service or use the prohibited AI systems. Therefore, it is important for the above operators to ensure that they do not place on the market or use a prohibited AI system. To do so, they will have to verify if the AI system in question falls under the prohibitions in Article 5.

Content of the first two prohibitions

14. **This call for input focuses on the prohibitions set out in subparagraphs (a) and (b) of Article 5, paragraph 1.** First of all, the AI Act prohibits AI systems that use manipulative or deceptive techniques (hereinafter: manipulative systems). In the remainder of this call for evidence, we will refer to this prohibition as 'prohibition A'. The Regulation defines this prohibition as follows:

> **Article 5, paragraph 1, subparagraph a ('prohibition A'):**
> "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 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;"

15. **Subsequently, the Regulation also prohibits AI systems that use exploitative practices (hereinafter: exploitative systems).** In the remainder of this call for input. We will refer to this prohibition as 'prohibition B'. This prohibition reads:

Article 5, paragraph 1, subparagraph b ('Prohibition B'):
"the placing on the market, the putting into service or the use of an AI system that exploits any of the vulnerabilities of a natural person or a specific group of persons due to their age, disability or a specific social or economic situation, with the objective, or the effect, of materially distorting the behaviour of that person or a person belonging to that group in a manner that causes or is reasonably likely to cause that person or another person significant harm;"

IV. Criteria and questions regarding the prohibitions

16. **In order to structure this call for input, separate criteria of the prohibitions have been set out in more detail in the next section.** These criteria are highlighted because they are important conditions for determining whether or not AI systems are covered by Prohibition A or Prohibition B. A brief explanation is provided for each criterion, based on the explanation provided by the legislator in the explanatory recitals to the AI Act and, in some cases, the AP's own interpretation. This is followed by some accompanying questions that you can use when providing your input.

Criterion 1 (prohibitions A and B): AI-enabled manipulative, deceptive and exploitative practices

17. **For both manipulative or deceptive systems (prohibition A) and exploitative systems (prohibition B) to be covered by the prohibitions, it is a requirement that it can be demonstrated that the practice is AI-enabled.** Many manipulative or deceptive practices are prohibited under consumer law in relation to unfair commercial practices. The prohibitions for such AI practices are complementary to the provisions of Directive 2005/29/EC of the European Parliament and of the Council, in particular those provisions which establish that unfair commercial practices leading to economic or financial harms to consumers are prohibited under all circumstances, irrespective of whether they are put in place through AI systems or otherwise.

Questions related to criterion 1

1. Can you give examples of systems that are AI-enabled and that (may) lead to manipulative or deceptive and/or exploitative practices?
2. Are you aware of AI systems where it is not sufficiently clear to you whether they lead to manipulative or deceptive and and/or exploitative practices? What do you need more clarity about? Can you further explain this?

Deployment of (i) subliminal, manipulative and deceptive techniques or (ii) exploitation of vulnerabilities

18. **The following sections contain separate descriptions of the components 'subliminal, manipulative and deceptive techniques' (prohibition A) and 'vulnerabilities' (prohibition B).** Those parts relate exclusively to Prohibition A or Prohibition B and therefore raise different questions.

Criterion 2 (prohibition A): Deployment of subliminal, manipulative and deceptive techniques

19. **Prohibition A applies to AI systems that deploy subliminal techniques, beyond a person's consciousness, or that are purposefully manipulative or deceptive.** The recital sets out that such AI systems can deploy subliminal components such as audio, image, video stimuli that persons cannot perceive or control. This concerns stimuli that go are beyond human perception, or other manipulative or deceptive techniques that subvert or impair person's autonomy, decision-making or free choice in ways that people are not consciously aware of those techniques or, where they are aware of them, can still be deceived or are unable

to control or resist them. The use of technology for machine-brain interfaces or virtual reality increases the possibility of such control.

Questions related to criterion 2

3. Do you know of examples of AI systems that use subliminal techniques?
4. Can you provide AI-specific examples of subliminal components such as audio, images or video stimuli that persons cannot perceive or control?

Criterion 3 (prohibition B): Exploitation of vulnerabilities

20. **Prohibition B applies to AI systems that exploit human vulnerabilities.** The Regulation further sets out that exploitative systems are prohibited if they exploit any of the vulnerabilities of a person or a specific group of persons due to their age – such as children and the elderly – disability (as defined in the European Accessibility Act (Directive (EU) 2019/882)) or specific social or economic situation that makes them more vulnerable to exploitation, such as persons living in extreme poverty or persons belonging to an ethnic or religious minority.

Questions related to criterion 3

5. Could you give examples of AI systems that, in your view, exploit vulnerabilities of a person or of a specific group of persons? Can you further explain this?
6. Are there any cases that may, in your view, may unjustifiably fall outside the scope of the prohibition?

Criterion 4 (prohibitions A and B): Significant harm

21. **A condition for the applicability of both prohibitions A and B is that the distortion of behaviour that causes or is reasonably likely to cause a person or persons significant harm.** This is in particular the case if sufficiently important adverse impacts on physical, psychological health or financial interests are likely to occur. This means that the prohibition goes beyond the sole protection of economic interests and actually is about providing protection against physical or psychological harm as well.

Questions related to criterion 4

7. Can you provide examples of systems that cause or are reasonably likely to cause significant harm to a person's physical or psychological health or financial interests? Do you have any questions or need specific clarifications where it concerns these prohibitions?

Criterion 5 (prohibitions A and B): With the objective, or the effect, of materially distorting behaviour

22. **It follows from the prohibitions that the placing on the market, putting into service or use of an AI system should have the objective or the effect of materially distorting behaviour of persons.** If the distortion of behaviour results from factors external to the AI system which are outside the control of the provider or

the deployer, it may not be possible to assume that the objective (intention) was to disrupt the behaviour. That is because these factors could not reasonably be foreseen.

23. **In the case of prohibition A, a material distortion of behaviour will occur if the objective or the effect is that an AI-system is to appreciably impair the ability of individuals to take an informed decision, which causes them to take a decision that they would not have taken otherwise.** The legislator has not further elaborated on what constitutes a distortion of behaviour in the event that exploitative systems are used (prohibition B). One can assume that the bar for applying this prohibition is lower in this respect, because it concerns the use of vulnerabilities of persons. For both prohibitions, a material distortion may apply both to individuals and to groups of persons.

Questions related to criterion 5

8. Do you know examples of AI systems that affect the ability of individuals to make decisions in such a way that the ability to make an informed decision may be distorted, which causes them to take a decision that would not have taken otherwise? Are there any open questions or specific clarifications needed on this topic?
9. Do you know examples of situations where the behaviour of one or more groups of persons is distorted by an AI system? Do you have any questions or need specific clarifications where it concerns these prohibitions?

Criterion 6 (prohibition A and B): Individual and group harm

24. **Prohibitions A and B apply if harm is caused or can reasonably likely be caused to the persons whose behaviour is distorted, or to other persons.** This can also include harm caused to a groups of persons.

Questions related to criterion 6

10. Do you know of situations where there is a material distortion of behaviour of a person or group of persons, leading to harm to persons other than that group and where AI systems play a role? Do you have any questions or need specific clarifications where it concerns these prohibitions?

Scope of the prohibitions

25. **The recitals of the Regulation also set out situations in which an AI system does not fall under the prohibitions.** Recital 29 of the Regulation clarifies that the prohibitions should not affect lawful practices in the context of medical treatment such as psychological treatment of a mental disease or physical rehabilitation, when those practices are carried out in accordance with the applicable law and medical standards, for example explicit consent of the individuals or their legal representatives. In addition, common and legitimate commercial practices, for example in the field of advertising, that comply with the applicable law should not, in themselves, be regarded as constituting harmful manipulative AI-enabled practices.

Questions related to criterion 7

11. Which medical treatments do you know that are AI-based and use subliminal techniques?
12. Are there possible other categories of purposes or applications of AI systems for which it is unclear whether they should be covered by the prohibitions? Do you have any questions or need specific clarifications where it concerns these prohibitions?

26. **Finally, it should be noted that this document does not cover all aspects of the prohibitions.** Interested parties are therefore invited to provide relevant input for the further explanation of Prohibition A and Prohibition B in addition to the questions asked in this call for input.

Final questions

13. Apart from the questions asked in this call for input, is there any relevant input that you would like to provide for the further explanation of Prohibition A and Prohibition B?
14. Would you prefer that we explicitly mention your organisation or group in our public summary and appreciation of this call for input, e.g. to address the examples and considerations you have provided?

Annex: overview of prohibitions from Article 5, paragraph 1 of the AI Act 2024/1689

Prohibition A: Certain manipulative AI systems

AI systems that deploy subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques, with the objective, or the effect of materially distorting the behaviour of a person or a group of persons by appreciably impairing their ability to make an informed decision, 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.

Prohibition B: Certain exploitative AI systems

AI systems that exploit any of the vulnerabilities of a natural person or a specific group of persons due to their age, disability or a specific social or economic situation, with the objective, or the effect, of materially distorting the behaviour of that person or a person belonging to that group in a manner that causes or is reasonably likely to cause that person or another person significant harm.

Prohibition C: Certain AI systems for social scoring

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 social score leading to either or both of the following:

- detrimental or unfavourable treatment of certain natural persons or groups of persons in social contexts that are unrelated to the contexts in which the data was originally generated or collected;
- detrimental or unfavourable treatment of certain natural persons or groups of persons that is unjustified or disproportionate to their social behaviour or its gravity.

Prohibition D: Certain AI systems for predictive policing

AI systems 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.

Prohibition E: Untargeted scraping of facial images

AI systems that create or expand facial recognition databases through the untargeted scraping of facial images from the internet or CCTV footage.

Prohibition F: Certain AI systems for emotion recognition in the workplace or in education

AI systems that infer emotions of a natural person in the areas of workplace and education institutions, except where the use of the AI system is intended to be put in place or into the market for medical or safety reasons.

Prohibition G: Certain AI systems for biometric categorisation of persons

AI systems for biometric categorisation 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.

Prohibition H: Certain AI systems for real-time remote biometric identification in publicly accessible spaces for purposes of law enforcement

AI-systems used for of 'real-time' remote biometric identification systems in publicly accessible spaces for the purposes of law enforcement, unless and in so far as such use is strictly necessary for one of the following objectives:

- the targeted search for specific victims of abduction, trafficking in human beings or sexual exploitation of human beings, as well as the search for missing persons;
- the prevention of a specific, substantial and imminent threat to the life or physical safety of natural persons or a genuine and present or genuine and foreseeable threat of a terrorist attack;
- the localisation or identification of a person suspected of having committed a criminal offence, for the purpose of conducting a criminal investigation or prosecution or executing a criminal penalty for offences referred to in Annex II and punishable in the Member State concerned by a custodial sentence or a detention order for a maximum period of at least four years.

Point (h) of the first subparagraph is without prejudice to Article 9 of Regulation (EU) 2016/679 for the processing of biometric data for purposes other than law enforcement.