AUTORITEIT
PERSOONSGEGEVENS

Manipulative, deceptive and exploitative AI systems

Summary of responses and next steps

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

December 2024
DCA-2024-04

Summary and next steps

This document provides a summary of responses and identifies follow-up steps in relation to the previous call for input regarding the first two prohibitions in Article 5 of the AI Act (see document DCA-2024-01). This concerns AI systems that use manipulative or deceptive techniques, as well as AI systems that exploit human vulnerabilities. Overall, the AP distils three main observations from the responses received.

Main observation 1 – Many respondents note that the practices to which the prohibitions relate are already regulated or prohibited by other horizontal or sector-specific laws and regulations.

The prohibitions in question are closely linked to consumer protection laws such as directives on unfair commercial practices, the Digital Services Act (DSA) and financial regulations. The practices also affect the lawful processing of personal data as regulated in, among others, the General Data Protection Regulation (GDPR). According to the AP, the prohibitions in the AI Act have a preventive nature (AI systems may not be developed or used) and therefore the prohitions complement the aforementioned laws. As such, the prohibitions provide an opportunity for more effective supervision of situations in which AI systems play a role in cases of deception, manipulation and exploitation of persons. It is important that these regulations are explained and applied as coherently as possible. This requires close cooperation and coordination between the Netherlands Authority for Consumers and Markets (ACM), Dutch Authority for the Financial Markets (AFM) and the AP.

Main observation 2 – Several respondents point to the additional possibilities that AI offers to deceive, manipulate or exploit people more convincingly.

A frequently cited example of this is the use of *dark patterns*, where interfaces are designed or organised in such a way that they deceive or manipulate people into making choices or decisions that they would not otherwise make. AI technology can reinforce manipulative or deceptive patterns by, for example, better responding to a person's characteristics or vulnerabilities during the interaction with the user.

Main observation 3 – Respondents note that the prohibitions are not elaborated in standards and therefore need to be further clarified in guidelines.

The prohibitions in the AI Act are new and contain connections and concepts that are operationalised only to a limited extent in the AI Act itself. This means there are also a number of topics that require further clarification, as is noticeable in the input from respondents. This concerns, for example, the extent to which there must be a link between the AI system and the deception or manipulation; the scope of the concept of harm; when *significant* harm has been caused; what exactly can be understood as a subliminal or deceptive technique; what types of vulnerabilities an AI system can exploit and whether a behavioural distortion can also occur gradually.

The AP uses these insights to support the preparation of Dutch supervisory authorities to supervise the prohibitions and will further incorporate the input received in the explanation of the prohibitions in the coming months. This involves cooperation and coordination with other supervisory authorities, including in the AI and Algorithm Chamber of the Digital Regulation Cooperation Platform (SDT). In addition, the input provides a basis for the AP's contribution from practice in the discussion on the guidelines on the prohibitions that the European Commission is currently working on.

I. Background

AI Act & prohibited AI

1. **The AI Act (2024/1689) entered into force on 1 August 2024.** This act sets out rules on the development and use of artificial intelligence (AI) in the EU. The starting point of the AI Act is that there are many useful applications of AI, but that the technology also entails risks that need to be managed. Some AI systems with unacceptable risk will be prohibited.

Call for input

2. **On 27 September 2024, the AP issued a call for input regarding the first two prohibitions in Article 5 of the AI Act.** This concerns AI systems that use manipulative or deceptive techniques (hereinafter referred to as prohibition A), as well as AI systems that exploit human vulnerabilities (hereinafter referred to as prohibition B).

3. **The call for input aims to gather information and insights from stakeholders (citizens, governments, companies and other organisations) and organisations that represent them.** The AP aims to collect information to provide a basis for preparing further interpretation of the prohibitions in the AI Act. The call for input is also part of a series of calls on the various prohibitions in the AI Act. More information about the AI Act, the prohibitions and the role of the AP in supervision can be found on the website of the AP and in the previously published call for input.

4. **Between 27 September and 17 November 2024, the AP received fifteen responses from various organisations and individuals with diverse backgrounds, including academics.** The AP is pleased that the call has reached citizens, companies, researchers and social organisations and that they have wanted to contribute by submitting input.

5. **Following the call for input, this summary document was prepared.** All input has been reviewed by the AP and the main points have been included in this document. The document refers only in generic terms to the submitters of the input received. In a limited number of places in this document, the input is also appreciated by the AP. This has been done where appropriate and where it could be traced back directly to the provisions in the AI Act or where an (additional) point of view could be provided. In doing so, the AP aims to contribute to the discussion on the interpretation of the prohibitions. With this document, the AP does not intend to provide an explanation of the prohibitions or related laws.

AI Act compliance supervision

6. **The AP made this call for input in its role as coordinating supervisor of algorithms and AI.** Within the AP, these tasks have been assigned to the Department for the Coordination of Algorithmic Oversight (DCA). The call for input is an extension of the preparatory work being done for the supervision of prohibited AI systems under the AI Act. The government is currently working on the formal designation of national supervisors for the AI Act. The AP (from the Department for the Coordination of Algorithmic Oversight) and the Dutch Authority for Digital Infrastructure (RDI) have issued an advice on this matter in collaboration and coordination with other supervisory authorities. It has been recommended, among other things, that the AP be designated as the market surveillance authority for most of the prohibitions in Article 5.

II. Summary and evaluation of submitted input

General points

Relationship with other laws and regulations

Various respondents note that the practices to which the prohibitions in question relate are in some cases already regulated or prohibited by other horizontal or sector-specific laws and regulations.

7. **Examples include the rules in consumer law, for example, when it concerns unfair commercial practices or *dark patterns* (including Article 25 DSA), or (specific) financial regulations that protect consumers against deception and manipulation.** The AP points out that several guidelines are already available on these rules, such as the Guidelines for the Protection of Online Consumers of the Netherlands Authority for Consumers & Markets (ACM) or the Policy Rules of the Dutch Authority for the Financial Markets (AFM) regarding the provision of information in the financial sector. Also, processing of personal data that underlies prohibited practices is regulated by the GDPR and processing of personal data for the purposes of the prohibited practices will also be unlawful.
8. **According to the AP, what is special about the prohibitions in the AI Act is that, in cases where manipulation, deception and exploitation involves an AI system, they relate to both the development and the use of such systems.** Therefore, both providers and users could be held liable. This will enable action to be taken against harmful practices involving AI at various points in the AI value chain. Furthermore, the prohibitions concern many different types of harm. In addition, the prohibitions in question also relate to the harm of *groups* of persons. This may make it less complicated to take action against prohibited AI practices than under other legal frameworks, which, for example, do not target providers, relate to psychological harm or concern groups of persons. That is why the prohibitions – in addition to, and interpreted in conjunction with, other rules governing similar practices – provide a means to supervise AI systems more effectively.
9. **The alignment of and between regulations is important.** According to the AP, the prohibitions should be interpreted and applied as much as possible in accordance with, and in conjunction with, existing (European) regulations in the areas of inter alia consumer protection and financial services, the DSA and the GDPR.
10. **In doing so, supervision of the prohibitions in the AI Act should focus primarily on practices in which AI technology plays an important role.** This requires close cooperation and coordination between the supervisory authorities, including in particular the ACM and AFM. The authorities are already making efforts to achieve this.

Relationship to high-risk uses and other prohibitions

Respondents note that the prohibitions also relate to high-risk applications and other prohibitions, and that clarity should be provided on this.

11. **Respondents raise the question whether the exception for AI systems that do not pose a significant risk to the health, safety or fundamental rights of natural persons (Article 6, third paragraph of the AI Act), also applies to the prohibitions.** Such systems would then not considered to be high-risk. According to the AP, this exception only applies to AI systems that would otherwise be considered high-risk AI systems (Chapter III of the AI Act). This exception therefore does not apply to the prohibitions in the AI Act.

12. **Several respondents also point to the link with the banned and high-risk AI systems for emotion recognition in relation to the prohibitions delineated upon in this summary.** This particularly concerns the possibilities that emotion recognition offers to better respond to people's vulnerabilities or to manipulate or deceive them more effectively.

Involving professionals

A respondent noted that developments in AI technology are occurring rapidly and that at the same time, there is a need for clarity on the application of the rules and best practices.

13. **AI professionals can contribute to clarification.** The AP considers this dialogue with stakeholders to be important so that the AP, together with other supervisory authorities, can contribute to the explanation and guidance on the AI Act. The respondent also emphasises the importance of education and practical training by the organisation. According to the AP, this also ties in with the subject of AI literacy (Article 4 of the AI Act) and the steps that organisations developing or using AI systems are taking to promote AI literacy.

Specific points

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

Several respondents indicated that manipulative, deceptive or exploitative practices are increasingly enabled by AI and that the use of AI increases the risks associated with these practices. According to respondents, the use of AI, in combination with the use of available personal data, makes it possible to approach and interact with people when they are at their weakest or most vulnerable. Because of their adaptability, AI systems are also able to respond well to a person's circumstances or vulnerabilities. Clarification seems particularly needed on the issue of the link between the manipulative, deceptive or exploitative practice and the AI system.

14. **Respondents often note the additional possibilities offered by AI to use *dark patterns*, where interfaces are designed or organised to deceive or manipulate people into making choices or decisions that they would not otherwise make.** For example, AI technology can be used to further tailor an interface and the information provided therein to a person's characteristics or vulnerabilities, thereby reinforcing manipulative or deceptive patterns in the interface.

15. **Several respondents also point to the possibilities that generative AI offers to deceive or manipulate people more convincingly.** Respondents also point out the possibilities of using AI to spread or recommend misinformation or manipulative content, or to persuade people to change their voting behaviour, for example. An individual respondent also noted that the risk of deception, manipulation or exploitation is greater if AI technology is more deeply integrated into other applications compared to when the technology is superficially and visibly present in, for example, a chatbot. Many of these concerns also relate to the regulation of *General Purpose AI models* and the transparency obligations in Article 50 of the AI Act. The supervision of the ACM and the ACM Guidelines regarding the Digital Services Regulation are also relevant here, as they largely relate to online platforms and, for example, the dissemination of misinformation and disinformation.

16. **Some respondents indicate that the necessary link between an AI system and manipulation, deception and exploitation needs clarification.** According to respondents, the question is whether and when practices involving manipulation, deception or exploitation are related to the use of AI. One respondent

noted that this aspect of the prohibitions is crucial to the scope of the prohibitions. In light of that, this respondent notes that using *other simpler technologies* in combination with AI can also lead to manipulation, deception and exploitation. According to the AP, this shows how important the explanation of the term 'AI system' is for the applicability of the prohibitions. According to the AP, this also underlines the importance of choosing an integral approach in the future interpretation of this definition with regard to practices in which AI technology is used. This in order to address the risks of the use of, for example, deceptive AI systems.

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

Respondents indicate that – in the framework of prohibition A – there is still a great deal of uncertainty about what can be understood by subliminal techniques of which people are not aware, or by manipulative or deceptive techniques. They argue that it is particularly necessary to clarify what such techniques entail. The extent to which manipulation or deception must have been intended also requires clarification.

17. **In the framework of the questions about subliminal techniques, several respondents mention examples such as systems that use imperceptible pitches or images that are displayed very quickly.** If such techniques can indeed lead to behavioural distortion, then according to the AP, it should be clarified whether these indeed constitute subliminal techniques within the meaning of the AI Act. Respondents also again mentioned the use of dark patterns in their answers regarding subliminal techniques.

18. **One respondent notes that further clarification should be provided as to how this prohibition can refer to *purposefully* manipulative or deceptive techniques.** Simultaneously, the section in the legal provision on the prohibition about the 'distortion of behaviour' refers to a distortion that is the objective *or* the *effect* of the use of the AI system. According to this respondent, it should therefore be possible for manipulative or deceptive techniques to also fall under the prohibition if they only have the effect (and therefore do not *necessarily* have the purpose) of materially distorting the behaviour of individuals. According to the AP, how this criterion, in particular the 'intention' of the use of manipulative or deceptive techniques, should be interpreted is indeed a point of concern.

Criterion 3 (prohibition B): Exploitation of vulnerabilities

In the framework of prohibition B, respondents note that AI technologies make it increasingly easier to exploit human vulnerabilities. According to respondents, it is also necessary to clarify what kind of vulnerabilities included in this prohibition and whether this may also involve the use of proxies that indicate a person's vulnerabilities.

19. **According to various respondents, it is unclear what exactly is meant by exploiting vulnerabilities.** One respondent indicated that it would be helpful if there were guidelines to assess whether vulnerabilities are being exploited. According to one respondent, the use of AI and the (personal) data processed in the process makes it possible to gain more insight into a person's circumstances, such as their financial circumstances. This would make it easier to exploit vulnerabilities associated with it. Another respondent notes that large data processing and the use of AI also make it possible to exploit people's vulnerabilities in less visible and subtle ways. In such cases, proxy data is used that does not directly indicate the presence of a vulnerability, but is an indirect indicator of it. According to the AP, this recognises the fact that the availability of more (personal) data and the increased possibilities to process this data with AI technologies only increases the risks of exploitation of vulnerable people. However, according to the AP, the interpretation of this prohibition ultimately depends on whether the AI system uses vulnerabilities that, as described in the AI Act, are the result of a person's age, disability or a specific social or economic

situation. So in this assessment, it is not necessary per se to assess how and on the basis of which (personal) data the AI system determines that there is a vulnerability.

20. **Lastly, respondents also believe that AI can be used to protect or support vulnerable people.** For example, one respondent indicates that AI can be used to provide timely information or signalling to vulnerable users of certain essential services, which is sometimes also a legal requirement imposed on parties. In addition, according to respondents, the use of AI technology also offers opportunities to, for example, physically, sensorily or cognitively support persons with disabilities by means of assistive AI technologies.

Criterion 4 (prohibitions A and B): Significant harm

Respondents indicate that it may be complicated (in certain cases) to determine what kind of harm may be caused in the event of a distortion of behaviour, and when this harm is significant.

21. **According to one respondent, significant harm, as described in the recitals of the AI Act as 'in particular having sufficiently important adverse impacts on physical, psychological health or financial interests', can take many different forms.** This respondent also indicates that the recital provides a non-exhaustive list of the types of adverse impacts due to harm.

22. **Respondents do not have a consistent picture of the scope of the concept of harm.** According to one respondent, the concept of harm would, for example, include social harm, e.g. as a result of disinformation, while another respondent indicates that this cannot be the case. According to the respondents, it is important that clarification is provided about which types of harm the prohibitions relate to.

23. **It is also noted by a respondent that a causal link must be established between the harm and the distortion of behaviour and that this is an element of this prohibition that requires clarification.** One respondent suggested that it could be helpful to establish certain 'presumptions' for determining this harm and to clarify which steps in thought are required when determining the causal link.

24. **One respondent indicates that in order to determine whether significant harm has occurred, the 'relativity' of harm must be taken into account.** According to the respondent, some harm for the same application of AI might not be significant to one person or group of persons, yet might be significant for another. That is why a flexible approach should be assumed when assessing the significance of harm. Some respondents note that it could be particularly complicated to assess whether significant harm has occurred if it is intangible. All in all, there is a need among respondents for more clarity about the 'significance' of the harm.

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

From the responses of various respondents, it follows that the criterion of 'the material distortion of behaviour' raises questions.

25. **A number of respondents indicate that the criterion of 'materially distorting behaviour' can also be found in consumer law.** According to some respondents, this concept has already been operationalised. This concept is also linked to the concept of the 'average consumer'. According to one respondent, the link with consumer law is clearer in prohibition A, where it is also explained that the criterion concerns appreciably impairing the ability to make an informed decision, thereby causing the persons to take a decision that they would not have otherwise taken. One respondent does indicate that this prohibition has a different rationale than provisions in consumer law, in which this concept is already used. That is why this criterion

should not be interpreted in the same way as in consumer law. According to the AP, the concept of 'the material distortion of behaviour' should be interpreted as much as possible in conjunction with existing laws. Yet the AP notes that the prohibitions in the AI Act also apply to situations that are not related to the purchase of products or services and the role of people as consumers. As a result, this prohibition likely requires additional operationalisation.

26. **It is also important that this criterion applies to people that are 'appreciably' impaired in their ability to make an informed decision.** The question arises from the input from respondents whether the prohibition only applies if a clear decision is made, such as purchasing a service or buying a product, or whether the prohibition also applies to situations in which the influence leads to a gradual change in behaviour or change in mood that is harmful or has harmful consequences for, for example, someone's health or well-being. According to the AP, clarification also appears to be necessary with regard to temporal properties of possible distortions of behaviour.

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

According to several respondents, it is important to emphasise that both prohibitions can involve harm suffered by both individuals and groups of persons.

27. **According to one respondent, it will have to be explained how different types of harm relate to the individual or to multiple persons.** One respondent indicated that it is particularly important to also draw attention to the harm to groups of persons. According to another respondent, this explanation should also take into account harm that 'other' persons may suffer whose behaviour has not been significantly distorted. One respondent indicated that there would be a difference in the actors who could suffer harm per prohibition. According to this respondent, the harm under prohibition B only relates to an individual or individual persons, while under prohibition A, it can also relate to group harm. According to the AP, a textual interpretation of both prohibitions seems to support that reading, but more clarity and certainty is desirable in this area also.

Scope of the prohibitions

According to respondents, there should be more clarity on the cases in which an AI system should not fall under the provisions on the prohibitions.

28. **One respondent indicated that this clarification should focus on the section on common and legitimate commercial practices, with this respondent emphasising that a *common* commercial practice is necessarily a legitimate one.** Other respondents indicate that it would be helpful to provide more insight into the types of legitimate practices in the context of medical treatment that should not be affected by the prohibitions in this act.

III. Follow-up

29. **The publication of this summary document concludes the information gathering process through the call for input on the prohibitions on manipulative, deceptive and exploitative AI systems**. The submitted input leads to a number of key points, which are relevant to the interpretation that will be given to the prohibitions in the AI Act:

    a. **Both prohibitions contain many terms and concepts that are not (or only to a very limited extent) explained in the AI Act.** Interpretation through (European) guidelines, as well as further clarification with specific examples, are crucial to ensure that these prohibitions offer protection to citizens and that legal security is provided to market parties. This concerns, for example, the scope of the AI system definition, the techniques to manipulate or deceive, the use of vulnerabilities and the harm that can also be suffered by groups.
    b. **In particular, the prohibition regarding manipulative and deceptive AI (Article 5, first paragraph, point (a) of the AI Act) relates to laws in the field of consumer protection, financial services and to the DSA**. Clarification is needed on how these prohibitions relate to concepts in those laws, given the independent nature and rationale of the prohibitions in the AI Act. Here, for example, it concerns the meaning of 'significant harm' or the material distortion of behaviour. In this context, many of the practices mentioned, such as the use of dark patterns, require special attention.
    c. **The prohibitions also relate to other rules in the AI Act, which will sometimes also be supervised upon at EU level.** This relation and the scope of these provisions should be clarified by the market surveillance authorities and the AI Office. Here it concerns the relationship of the prohibitions with e.g. transparency obligations (Article 50), other prohibitions or high-risk applications (such as emotion recognition) and the rules for general purpose AI models.

30. **The AP is working on the preparation of the supervision on the prohibitions in the Netherlands and, in the coming months, will further use the input received for the preparation of information and explanation on the prohibitions.**. The knowledge gained is shared with Dutch supervisory authorities in the Digital Regulation Cooperation Platform (SDT), and the AI and Algorithm Chamber (AAK) of the SDT will discuss which follow-up actions are needed to clarify the prohibitions in the AI Act. The above approach is in line with the vision of supervisory authorities to strive for a harmonised interpretation and application of rules that can be simultaneously applied to prohibited AI practices.

31. **The AP will also use the insights gathered from practice as a basis for contributing to the discussion on the guidelines on the prohibitions. The European Commission intends to publish the first guidelines in early 2025.** To this end, it has launched a consultation for stakeholders, including AI system providers, companies, national authorities, academics, research institutions and civil society. More information on these guidelines and consultation can be found on the European Commission website.