WilhelmGilliéron
AVOCATS

Auteur: Philippe Gilliéron | Le : 15 February 2024

The EU AI Act – 2 – Classification : prohibited practices and general purpose AI models

The EU AI Act sets requirements depending upon the intensity and scope of the risks that AI systems can generate, in light of the seven principles that underly the Regulation and that we mentioned in our previous paper, in particular, but not only, with regards to fundamental rights.

This risk-based approach led the Commission to make a distinction between (i) a certain number of activities that are deemed unacceptable and that have to be prohibited, (ii) systems considered high risk that are at the core of the Regulation, as well as (iii) transparency obligations for certain AI models and systems, notably general purpose AI (GPAI) models.

In this second paper of our series devoted to the EU AI Act, we shall focus on the prohibited practices (Title II) and the transparency obligations for providers and deployers of certain AI systems and GPAI models (Title IV)

I. Prohibited practices (Title II, Art. 5)

Shall be prohibited the placing on the market, respectively putting into service or use of the following AI systems:

a) AI-enable manipulative techniques

Systems using subliminal techniques (such as audio, image or video stimuli that are beyond human perception) or any manipulative of distorting techniques that are meant to persuade people to engage in unwanted behaviors that these persons are not consciously aware of should be prohibited. The same goes with regards to such systems that exploit vulnerabilities from a person of a specific group (such as children, elderly people, ones suffering from disabilities or poor economic or social conditions) with the objective or the effect to distort their behavior.

In both instances, such prohibition however only exists if the use of the AI system at stake cause or is likely to cause a significant harm to the individual, including harms that may be accumulated over time.

Although intent will in most instances be present, intent would not be required; all that matters is the objective impact of such AI systems.

This prohibition should also be red in light of the provisions contained in Directive 2005/29/EC related to unfair commercial practices, bearing in mind that, according to the preamble of the Regulation, common and legitimate practices notably in the field of advertising that comply with the applicable law should not in themselves be regarded as constituting harmful manipulative practices.

While advertising is by definition meant to induce the recipients to a certain behavior, one may wonder whether the use of subliminal techniques should not, as a result of Art. 5, be prohibited *per se* in that industry as well. The way the notion of "significant harm" will be construed will certainly play a key role: may the purchase resulting from the use of subliminal techniques in advertisement (or in the gaming industry to induce in-game purchase) be considered a significant harm? Should it depend upon the amount at stake? In a world where the asymmetry of information is always wider, transparency in my view becomes more important than ever and would lead me to answer in a positive way. Not sure, though, that this will be the outcome of the construction of "significant harm". Wait and see.

The use of such systems in the context of medical treatment, such as mental disorder or physical rehabilitation, does however not fall under that prohibition; provided, obviously, that these practices are carried out in line with the applicable medical regulatory framework.

b) Biometric categorization and social scoring

These prohibitions refer to the use of biometric data, *i.e.* the collection of personal data resulting from specific technical processing relating to the physical (facial), physiological or behavioral characteristics of an individual, to achieve certain goals that are considered unethical and, on that regard, illegal under the Regulation.

Is prohibited the use of such data to infer individuals' political opinions, trade union membership, religious or philosophical beliefs, race, sex life or sexual orientation, all considered as special categories of data within the meaning of Art. 9 GDPR which as a result, deserves special protection (without mentioning the fact that such inferences or correlations based upon biometric data may be questionable from a scientific standpoint on several accounts).

Is further prohibited the use of such data for social scoring, *i.e.* evaluate or classify people over a certain period of time based upon their social behavior inferred from multiple data points; provided, however, that such social scoring leads to either unfavorable treatment (i) in social contexts unrelated to the contexts in which such data has originally been generated or collected or (ii) unjustified or disproportionate to their social behavior or its gravity. As a result, and at least in my view, this means that any profiling activity falling under that provision would be prohibited, no matter whether it is carried out in line with the GDPR (which in any case remains doubtful when such processing violates the right to dignity and non-discrimination).

c) Real-time remote identification systems

The use of real-time remote identification systems in publicly accessible spaces for the purpose of law enforcement has been heavily debated during the negotiations, by fear of the risk of skidding and loss of control.

Ultimately, such use *may* be allowed by Member States (but it is therefore up to each Member State to decide) in the following narrow circumstances and stringent formal requirements that can be summarized as follows:

(i) Circumstances

The use of real-time remote biometric identification system for law enforcement may only be used for the following reasons:

- Targeted search for specific victims of abduction, trafficking and sexual exploitation of human beings as well as missing people;
- Substantial and imminent threat to the life or physical safety of natural persons or of a terrorist attack which, according to the preamble, would include serious disruption of a critical infrastructure within the meaning of Art. 2 (a) of Directive 2008/114/EC (where the question of knowing what is considered substantial and imminent may occur);
- Perpetrator or suspect of having committed one of the crimes listed in Annex IIa (which notably includes child pornography, trafficking of narcotic drugs or weapons, murder or grievous bodily injury, kidnapping, rape, environmental crime, etc.); provided, however, that those crimes should be punishable in the relevant Member State by at least 4 years.

Such use by law enforcement is only allowed in publicly available spaces. These spaces can be publicly or privately owned. What matters is that they are accessible to an indefinite number of people, regardless of whether certain conditions for access may apply

(such as, for instance a ticket for an event, to enter a fitness or a swimming pool). Would access granted through a badge in a privately held company still be considered as a publicly available space if thousands of people can get access to it, including guests, etc.? I would tend to answer in the positive, but questions are likely to occur as to what is considered a "publicly available space", in particular when it is privately owned.

In these circumstances, the EU AI Act will be considered as a *lex specialis* in respect of Art. 10 GDPR and will be considered as the legal basis for the processing of personal data under Art. 8 GDPR. Any other use of a biometric identification system, whether in real time or not, including by authorities, shall always be subject to the requirements set forth in Art. 9 and 10 GDPR (bearing in mind that several data protection authorities have already banned such use).

(ii) Formal requirements

The use of such systems in the above-mentioned circumstances will always be subject to:

- a prior authorization granted by a judicial authority based upon a reasoned request to be addressed in accordance with the rules that will have to be laid down for such request in each Member State accepting the use of such systems.
- In case of duly justified urgency, the use of such systems may start without such authorization, which shall however be requested at the latest within 24 hours; should the authorization not be granted, the use should be stopped immediately and all output deleted.
- Authorization should only be granted as necessary both in terms of time, geography and scope. Such decision should be based upon the nature of the situation (seriousness, probability and scale of the harm caused in the absence of such use) compared with the consequences of the use of such systems for the rights and freedoms of the persons at stake.
- The outcome of such assessment will require the authority to carry out a fundamental rights impact assessment as set forth in Art. 29a, based upon a template that the AI Office will have developed (automated tool).
- The system will have to be registered by the authority in the database set forth in Art. 51.
- The relevant market surveillance authority (pursuant to Regulation 2019/1020) and national data protection authority should be notified of each use. These authorities should then submit a yearly report to the Commission on the use of "real-time biometric identification systems" (based upon a template to be provided by the Commission).

d) Various

The following AI systems that are hard to categorize in one of the above are also prohibited:

- Risk assessments in order to assess the risk of a natural person to commit a criminal offense, based solely on the profiling of a natural person or on assessing their personality traits; provided, however, that such use should be allowed when it is already based on objective and verifiable facts directly linked to the criminal;
- AI systems that create of expand facial recognition databases through the untargeted scraping of facial images from the internet of CCTV footage (as a reaction in light of the Clearview case);
- AI systems that infer emotions of a natural person in the areas of workplace and education institutions (with the exception of such system used for medical or safety reasons).

II. General Purpose AI Models (Title IV)

a) Preliminary remarks

The advent of GPTs in 2023 is one of the reasons which explains the delay in the adoption of the EU AI Act. While several Member States were keen on ensuring that a certain level of control upon these systems would find its way into the EU AI Act, others such as France and Germany were more reluctant.

Ultimately, Member States have reached a compromise and Title IV now provides for certain transparency obligations for providers and users of such systems in its Art. 51 *et seq.* In a provision mirroring Art. 27 GDPR, the EU AI Act requires providers of GPAIs established outside of the Union to appoint an authorized representative to act as a contact point for the authorities.

Similarly to the notion of AI systems (see our latest paper), the Commission defines general purpose AI models (GPAI) based upon their functional characteristics, namely models that display significant generality and that are capable to perform a wide range of distinct tasks regardless of the way the model is placed on the market and that can be integrated into a variety of downstream systems or applications, including through libraries, APIs or as direct download. According to the preamble, models with at least one billion parameters should be considered as displaying such significant generality.

While focusing in this Title IV on GPAIs, the EU AI Act first addresses in its Art. 52 chatbots and emotion recognition biometric

categorization systems. In both instances, providers should inform the concerned natural persons that they are interacting, respectively being subject to such systems, and ensure that their data, notably with regards to emotion recognition or biometric systems, is processed in line with the application data protection legal framework.

The EU AI Act then turns to GPAIs. It makes a distinction between general purpose AI models (GPAI) in general, and the ones with systemic risk, that are subject to additional obligations.

b) General Purpose AI Models (GPAIs) in general

Generally speaking, providers of GPAIs have to ensure that:

- Outputs are marked in a machine-readable format and detectable as artificially generated or manipulated; provided, however, that such obligation is not imposed upon systems means to have an assistive function for text editing that do not substantially alter the input data.

  On February 6, 2024, Meta for example announced that AI generated content, including from industry partners such as Google, OpenAI, Microsoft, Adobe, Midjourney, and Shutterstock would be labelled.

- When the output leads to a deep fake, *i.e.* an AI generated or manipulated image, audio or video content that resembles existing persons, objects, places or other entities or events that would falsely appear to anyone to be authentic of truthful, providers should disclose that the content has been artificially generated or manipulated (which we understand to be a visible and perceptible notice for users that go beyond a watermarking in a machine-readable format).

- Interestingly, when the GenAI generates text meant to be publish for informing the public on matters of public interest, the disclosure that the text has been artificially generated does not apply where the AI-generated content has undergone a process of human review or editorial control and where a natural or legal person holds editorial responsibility for the publication of the content.

  Would this mean that the use of AI-generated text by media for their publication would be exempted from such a transparency requirement?

  All the above notices should be provided in a clear and distinguishable manner at the latest at the time of the first interaction.

- Taking into account the fact that these models may be integrated or form part of an AI system, the EU AI Act also provides for the obligation to draw up and keep up to date technical documentation of the model, including its training, testing process and the results of its evaluation, as well as information and documentation enabling providers of AI systems to understand the capabilities and limitation of the GPAI at stake. Further elements to be provided will be defined in forthcoming Annexes to the Act.

The above requirements will not apply to AI models that are made accessible under a free and open license such as, for instance, Llama-2, launched by Meta, and whose parameters including the weights and information on architecture model are made available.

Even open source licensed GPAIs will however have to comply with the following requirements:

- Put in place a policy to respect copyright law (notably so as to be able to detect holders having opted out in accordance with the data mining provision set forth in Art. 4 of the Directive 2019/790). The preamble provides that any provider placing a GPAI on the EU market should comply with this obligation, regardless of the jurisdiction in which the copyright-relevant acts underpinning the training of the GPAI takes place.
- Sufficiently detailed summary about the content (i.e. data) used for training (based upon a template to be provided by the AI Office). This summary should be generally comprehensible rather than technically detailed to enable copyright holders to exercise their rights, for instance by listing the main data collections or sets that went into the training model (such as large private database, data archives, etc.).

Unless the GPAI at stake presents a systemic risk, in which case no exception to above obligations shall apply, none of these above obligations are imposed when:

- an own model is used for internal processes that are not essential for providing a product or a service to third parties and the rights of natural persons are not affected.

- These models are used before release on the market for research, development and prototyping activities.
- These models are authorized by law to detect, prevent, investigate or prosecute criminal offense (with the limitations set out by the prohibited practices' section).

It is to be pointed out that the use of such a model for internal processes may not be considered as falling under an exception of the Union copyright law, so that the exemption for such uses to provide a copyright policy is, from the first look of it, questionable.

c) GPAIs with systemic risks

(i) Classification

A systemic risk is defined as having a significant impact on the internal market due to its reach, and with actual or reasonably foreseeable negative effects on public health, safety, public security fundamental rights or the society as a whole, and which can be propagated at scale across the value chain.

A GPAI model shall be classified as displaying systemic risks either if it has high impact capabilities (*i.e.* capabilities that match or exceed the most advanced GPAI) or based on a decision of the Commission taken ex officio or following an alert by the scientific panel (taking into account different criteria such as quality or size of the training dataset, number of business and end users, input and output modalities, degree of autonomy and scalability or the tools it has access to).

Is presumed to have high impact capabilities a model whose cumulative amount of compute used for its training measured in floating point operations (FLOPs) is greater than 10^25, a threshold that may evolve over time. As an example, it is estimated that ChatGPT was trained on 10^24 FLOPs, meaning that any models significantly more powerful than GPT-3.5 will be considered to bear systemic risk.

Assuming the provider considers, based upon its own assessment, to meet the high impact capabilities requirement, it will have to notify the AI Office at the latest two weeks after the requirements are met or after having found out that the FLOP threshold in particular will be met, together with the relevant information. The provider should however be entitled to demonstrate that, notwithstanding such threshold, the GPAI at stake does not present a systemic risk due to its specific characteristics. The Commission may however decide to reject those arguments and consider such GPAI to be of systemic risk. In such cases, the provider may request reassessment of its model every six months based upon objective, concrete and new reasons.

The Commission shall publish a list of GPAIs with systemic risks and keep it up to date.

(ii) Requirements

Taking into account the specific risks represented by such GPAIs models, their providers are subject to additional obligations meant to identify and mitigate those risks and ensure an adequate level of cybersecurity protection, regardless of whether such model is provided on a standalone basis or is embedded in an AI system. These providers shall:

- Perform model evaluation in accordance with standardized protocols and tools, including adversarial testing of the model with a view to identify and mitigate systemic risk;
- Continuously assess and mitigate possible systemic risks, including their sources that may stem from such models;
- Keep track of, document and report without undue delay to the AI Office and, as appropriate, to national competent authorities, serious incidents (*i.e.* which leads to the death or serious injury or serious damage to property or the environment, serious and irreversible disruption of the management and operation of critical infrastructure [understood as an asset, facility, equipment, network or a system which is necessary for the provision of an essential service as set out in Art. 2(4) of Directive 2022/2557], as well as a breach of obligations under Union law to protect fundamental rights) and their possible corrective measures.
- Ensure an adequate level of cybersecurity protection and the physical infrastructure of the model. Such level could be facilitated by securing model weights, algorithms, servers and datasets, such as through operational security measures for information security, specific cybersecurity policies, adequate technical and established solutions as well as cyber and physical access controls.

d) Code of practice (art. 52e)

While it is to be hoped that standards leading to a presumption of conformity will emerge overtime, the EU AI Act provides that the AI Office shall encourage and facilitate the drawing up of codes of practices at Union level, in particular with regards to the obligations applicable to GPAIs with systemic risks. The preamble provides that these codes should represent a central tool for the proper compliance with obligations foreseen under the Regulation, and that these providers should be able to rely on these Codes to demonstrate compliance.

Goal of these codes would notably be to ensure that (i) these obligations are kept up to date in the light of market and technological developments, (ii) type and nature of systemic risks and their sources are identified, as well as (iii) the measures, procedures and modalities for the assessment and management of systemic risks, including the documentation thereof.

It would then be up to the Commission to approve such a code or, alternatively, to provide common rules for the implementation of the obligations put upon the providers of GPAIs presenting systemic risks.

While the drafting of such Codes certainly should be encouraged, the idea of leaving it up to some extent to the providers of these Codes to draft them, for instance through Partnership on AI, may also be perceived as an acknowledgment that, taking into account the complexity and opacity of most of these models, the asymmetry of information makes it difficult for outsiders to rule upon those models. Although understandable, this obviously is regretful as one may fear that we will end up leave it up to the main stakeholders to set their own rules to a large extent.

In our third paper of this series, we shall focus on high-risk systems.

Source : https://www.wg-avocats.ch/en/actualites/intellectual-property/the-eu-ai-act-2-classification-prohibited-practices-and-general-purpose-ai-models/