cdt EUROPE

A Series on the EU AI Act

Pt. 3 – Freedom of Expression

May 2024

Authored by
***Laura Lazaro Cabrera**, Counsel and Director of the Equity and Data Programme, CDT Europe*

The use of AI systems has become commonplace in online communications. Recommender systems rely on AI to show individual users personalised feeds or recommended content, and synthetically generated images are increasingly circulated on social media. It is therefore unsurprising that the AI Act would have a role to play in safeguarding freedom of expression online and intersect with pre-existing EU legislation regulating online content governance, namely the Digital Services Act (DSA).

In this explainer, the third in our AI Act series, we reflect on the key changes introduced by the AI Act that will affect freedom of expression, its interaction with existing EU legislation, and key unanswered questions.

The AI Act impacts the way in which content is used or disseminated in two important ways. First, the Act prohibits some AI-fuelled practices that would result in the manipulation of individuals. Second, the Act imposes additional obligations on actors using AI to create synthetic content, including deepfakes. While freedom of expression is not singled out as a human right that the AI Act sets out to protect in connection with AI systems at large, the effect of some of the Act's provisions is to provide safeguards against some forms of AI-based manipulation known to hinder the right to hold opinions, a core element of the right to freedom of expression. However, the requirements for showing a violation of these provisions are stringent and may limit their effect.

The AI Act's Role in Outlawing Manipulative AI Systems

As we detailed in our earlier explainer, the AI Act adopts a risk-based approach that places AI systems in different risk-categories depending on their characteristics. One of the most well-known features of the Act is that it outlaws eight types of AI systems that it deems to pose unacceptable risks. Two of the prohibitions ostensibly seek to preserve freedom of opinion and will likely impact AI deployed in online platforms: the prohibition on AI systems' exploiting persons' vulnerabilities as recognised by the Act, and the prohibition of AI systems manipulating individuals.

Article 5(1)(a) prohibits AI systems that deploy subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques, with the object or effect of materially distorting a person's behaviour by "appreciably impairing a person's ability to make an informed decision, thereby causing a person to take a decision that that person would not have otherwise taken," in a way that could lead to significant harm. This provision targets at once manipulation and disinformation, bringing into scope AI systems that push people into making decisions they would not otherwise have made.

Although on its surface this provision would appear to be quite broad, the requirements to show a violation set a high hurdle. Instead of outlawing AI systems applying manipulative or deceptive techniques outright, the Act opts for an approach which requires a showing of i) a decision taken by a person that they would not have taken but for the AI system's operation, and ii) likelihood of the person incurring significant harm. The very nature of the techniques covered by the prohibition makes detection of their operation difficult. Once their existence is established, another hurdle is to isolate the effect of the AI system on a person from other factors likely to influence their behaviour, combined with the difficulty inherent in showing that any decision taken would not have been made *but for* the AI system. Lastly, a person would need to establish the risk of "significant harm," which the Act does not define.

Another prohibition relevant to freedom of opinion is contained in Article 5(1)(b), which is narrower in scope but presents similar challenges to Article 5(1)(a). It outlaws AI systems exploiting persons' vulnerabilities based on their age, disability, or specific social or economic situation, with the objective or effect of materially distorting the behaviour of a person in a way that could lead to significant harm. This provision could cover an AI system either directly interacting with a person – such as chatbots – or one delivering content recommendations. The prohibition is in itself noteworthy as it is prescriptive of the areas of vulnerability that may bring an AI system within the scope of the prohibition. The Act does not define the concept of "vulnerabilities" either on its own or in connection with each of the characteristics listed by the Act. Read broadly, the provision is likely to capture AI systems exploiting or intending to exploit individuals based on their youth or old age, difficult personal or financial circumstances, and disabilities resulting in cognitive impairments or challenges fuelled by disabilities regardless of their specific nature.

While the recognition of these key areas of vulnerability is welcome, the enforceability of this provision looks to be challenging in practice. In order for the prohibition to apply, either the AI system must (i) be intended to exploit vulnerabilities such that it would materially distort the targeted

person's behaviour in a way that would lead to significant harm, or (ii) actually and materially distort the targeted person's behaviour in a way that could lead to significant harm. For the same reasons articulated above, it will be a high hurdle for anyone to prove that either of the two scenarios described above will likely have occurred. In particular, regarding scenario (i), while intent may surely be presumed in some circumstances, the Act clearly states in its recital (29) that such intent may not be assumed where the distortion results from factors "external to the AI system which are outside the control of the provider or the deployer, namely factors that may not be reasonably foreseeable." In practice, a person seeking to establish that scenario (i) applies in relation to a specific AI system will face the onerous task of gathering information that may not be readily available and that may only become obvious through extensive testing of the AI system.

The AI Act's prohibitions, despite their attempts to limit the harmful manipulation of individuals online, are not a silver bullet against online disinformation – both of the prohibited practices which seek to protect freedom of opinion establish a high bar for anyone to meet. Because of this, their real impact will likely be unseen.

Interaction with the Digital Services Act

The article on prohibited AI systems has special relevance when read alongside the DSA. The DSA defines "illegal content" broadly to include content that is either itself illegal under existing law, or content that "applicable rules render illegal in view of the fact that it relates to illegal activities," irrespective of the subject-matter or nature of that law. The AI Act regulates the functioning or purpose of AI and not "content" as such; however, read in light of the DSA, it is entirely possible that content generated by or in furtherance of prohibited – read illegal – AI practices may also be illegal under the DSA and trigger related obligations.

As discussed in an earlier DSA explainer, the DSA creates new obligations for online platforms, ranging from ensuring the availability of mechanisms enabling the flagging of potentially illegal content (see the notice and action mechanism in Article 16) to abiding by the order of a competent authority to remove illegal content (see Article 9). These provisions are likely to enable better monitoring of the use of AI in online platforms by expanding the vectors for third parties – individual users as well as competent authorities foreseen by the DSA –- to seek removal of illegal content. However, it is unlikely that they will result in content related to prohibited AI practices being removed. Online platforms will be rightly wary of striking the right balance between the right to freedom of expression and the need to provide a safe online space, and therefore notices by individuals will likely only lead to the removal of content if it is manifestly illegal, a high hurdle to clear.

The DSA and the AI Act intersect in yet more ways. Both instruments seek to curtail manipulation of individuals, but address the issue differently: while the AI Act focusses on AI systems that operate in manipulative and deceptive ways, the DSA more narrowly focuses on dark patterns on online platforms, namely practices that materially distort or impair, either on purpose or in effect, the ability of recipients of the service to make autonomous and informed choices or decisions. The language used in both provisions is distinct: while the AI Act seeks to prohibit manipulative AI systems, the DSA uses comparatively lighter language by calling on providers not to design or use online

interfaces in a way that would deceive or manipulate users (Article 25). However, the threshold imposed in the DSA is much lower than that established by the AI Act in that the DSA's definition of manipulation does not require a showing of a decision by the user that they would not otherwise have made, or a showing of a likelihood of significant harm. In other words, the DSA has a broader conception of manipulation than the AI Act, but is of narrower application as it only applies to online platforms.

Both the DSA and AI Act also regulate profiling. The DSA bans the presentation of advertising based on profiling using special categories of data as defined by GDPR (Article 26(3)), as well as the presentation of advertising based on profiling to minors, regardless of the type of data used (Article 28 (2)). By contrast, the AI Act takes a top-level approach, deeming that any AI system deployed in any of the areas listed in Annex III (e.g., biometrics and law enforcement) *and* using profiling will be conclusively deemed to be high-risk. The starting point under the AI Act is therefore that AI systems using profiling are allowed, albeit subject in certain areas to the obligations created for providers and deployers of AI systems which fall into the 'high-risk' category. Further, some of the prohibitions on AI systems imply that profiling in narrow cases will be barred – such as the prohibition on exploiting vulnerabilities related to specific characteristics or the prohibition on social scoring. But such a bar would apply only when the profiling is accompanied by an intent to exploit vulnerabilities or has this effect in the case of Article 5(1)(b), or actually results in detrimental or unfavourable treatment in the case of Article 5(1)(c).

Here again, the DSA is more comprehensive as it clearly prohibits profiling-based advertising in specific circumstances. For example, the DSA prohibits targeting ads to minors based on profiling, while the AI Act prohibits AI systems exploiting vulnerabilities in relation to age but only if the system is intended to cause or actually causes a material distortion of behaviour likely to lead to significant harm. Unlike the AI Act, the DSA presumes harm from profiling-based advertising to minors.

Another intersection between the two laws concerns recommender systems. The DSA defines a recommender system as a fully or partially automated system used by an online platform to suggest or prioritise information to users. While recommender systems as such are not mentioned by the AI Act, many such systems plausibly fall within the Act's definition of AI system, namely a machine-based system designed to operate with varying levels of autonomy, that may exhibit adaptiveness after deployment.

However, it is unclear whether a recommender system would be considered a high-risk within the meaning of the Act. Annex III of the AI Act specifies the areas in which deployment of an AI system should presumptively lead to a "high-risk" categorisation, including systems used with the intent to influence the outcome of an election/referendum, or voting behaviour or natural persons in the exercise of their vote. Practically, an AI system being categorised as "high-risk" translates into an obligation for providers to seek a conformity assessment prior to placing the system on the market and for deployers to disclose and to ensure that AI systems directly interacting with natural persons are designed and developed in such a way that the concerned individuals are informed that they are indeed interacting with an AI. Going forward, an unanswered question is whether recommender systems suggesting content relevant to electoral processes would be considered to be high-risk under Annex III.

Even if they are high risk, the AI Act would not substantially add to the provisions already contained in the DSA. For example, the DSA already requires providers of online platforms using recommender systems to set out the main parameters used in their recommender systems, as well as presenting options for these parameters to be modified (Article 27); and imposes a requirement on providers of online platforms that they ensure they provide at least one option for each of their recommender systems which is not based on profiling (Article 28). While the AI Act does impose requirements on providers of high-risk AI systems, it does not offer the same degree of transparency and choice as the DSA.

The AI Act's Role in Content Regulation

As discussed previously, the AI Act's approach to AI systems is less prescriptive than the DSA's approach to online platforms – and where both instruments intersect, the DSA is more robust in terms of the conceptual ease of application of the obligations it imposes in relation to the operation of online platforms. Another key aspect where the AI Act intersects with online platform regulation is in relation to its approach to the different types of content generated by AI.

The AI Act recognises two distinct concepts: (i) synthetic audio, image, video, or text content, and (ii) deepfakes, which the Act defines as "AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places or other entities or events and would falsely appear to a person to be authentic or truthful." Each of these concepts invites different disclosure obligations, some of which are unique in the Act in that they do not apply to other AI systems, and differ from the obligations imposed for high-risk AI systems or GPAI models posing a systemic risk.

As a starting point, the AI Act calls on providers of AI systems intended to interact directly with natural persons to ensure that these systems are designed such that individuals interacting with them are aware that they are interacting with an AI, unless this would be obvious. In addition to this, there are two layers of disclosure requirements applicable to synthetic content: one at the level of the AI generating the content to flag that its outputs are synthetically generated, and another – applicable only in specific circumstances – at the level of the deployer using the AI. When the AI system or GPAI model produces any synthetic content, the Act creates an additional obligation for the providers to ensure that these outputs "are marked in a machine-readable format and detectable as artificially generated or manipulated" (Article 50(2)). If the AI system or model generates or manipulates text and is published with the purpose of informing the public on matters of public interest, or alternatively the content in question rises to the definition of deepfake, then an obligation is imposed on the *deployer* – in addition to the existing obligations on providers – to disclose that the content has been artificially generated or manipulated (Article 50(4)). The Act does not further elaborate on what that disclosure could look like.

The AI Act's choice of wording around both disclosure obligations applicable to synthetically-generated content and deepfakes is noteworthy, not least because it diverges from existing disclosure obligations applicable to online content under other laws. The Political Advertising Regulation, for example, flatly requires political advertisements to be accompanied by a statement that they are political advertisements, alongside other information which shall be made available in a "clear, salient and unambiguous way." More specifically, the Regulation calls on data controllers

to provide "meaningful information on the use of AI systems in the targeting or ad-delivery of the political advertising". By contrast, the AI Act simply requires for the artificially-generated or manipulated nature of the content to be "detectable," a concept which remains vague and falls short of requiring an explicit notice that would be clearly understandable to users. Further, the obligation on deployers to disclose that content has been synthetically generated does not contain any meaningful details as to the robustness and clarity of the disclosure. Despite the layered disclosure obligations, the AI Act's disclosure standards remain comparatively vague against the backdrop of existing EU legislation.

Notwithstanding the above, the Act quietly introduces a crucial exemption to the disclosure obligations around synthetic content. Crucially, the Act lifts the obligations applicable to providers and deployers to disclose the operation of an AI system, irrespective of whether it generates deep fakes, where the use is authorised by law to detect, prevent, investigate or prosecute criminal offences. In other words, the Act enables the undisclosed use of synthetic content for criminal investigations, so long as there is a basis in law to do so. This creates a new vector for the legal use of synthetic content by law enforcement, a surprising development in light of law enforcement's documented struggle with the criminal use of synthetic content.

Conclusion

The AI Act and the DSA intersect in various ways, but the standard of protection afforded by their provisions is markedly different, a fact made possible perhaps by the specificity and narrower scope of the DSA. However, even accounting for this fact, an assessment of both the AI Act and the DSA makes clear that the former does not go as far as the latter when it comes to protecting freedom of expression and opinion.

The AI Act's provisions regulating the providers and deployers of AI systems or models generating synthetic content or deepfakes are a regulatory novelty. However, the transparency and information requirements imposed for synthetic content seem to fall short of the standards existing elsewhere in EU legislation.

Find more from the CDT Europe team on the EU AI Act at *cdt.org*.

*The **Center for Democracy & Technology (CDT)** is the leading nonpartisan, nonprofit organization fighting to advance civil rights and civil liberties in the digital age. We shape technology policy, governance, and design with a focus on equity and democratic values. Established in 1994, CDT has been a trusted advocate for digital rights since the earliest days of the internet. The organization is headquartered in Washington, D.C. and has a Europe Office in Brussels, Belgium.*