REGULATING THE UNSEEN: THE EU AI ACT AND THE CHALLENGE OF SUBLIMINAL TECHNIQUES IN DISINFORMATION

Jaroslav Denemark*

Abstract: *This article examines the regulation of AI-enabled subliminal techniques used in disinformation under the EU Artificial Intelligence Act (AI Act). It explores how subliminal methods—such as deepfakes, psychographic microtargeting, and bot-driven amplification—challenge democratic discourse by exploiting psychological vulnerabilities beyond conscious awareness. Through a doctrinal analysis of the AI Act, the article assesses the prohibition of subliminal techniques under Article 5(1)(a), the ambiguous notion of "significant harm," the absence of high-risk classification for subliminal systems, and the limited obligations imposed on systemic-risk general-purpose AI models. Particular attention is paid to the transparency rules of Article 50, which are undermined by broad exemptions and the limited scope of responsibilities for very large online platforms. The findings reveal that while the AI Act acknowledges the risks posed by AI-driven disinformation, its current framework only partially addresses subliminal manipulation. The article concludes that significant regulatory gaps remain, especially in ensuring effective protection against covert AI-driven persuasion strategies in the information space.*

Keywords: *Artificial Intelligence Act; subliminal techniques; disinformation; deepfakes; psychographic microtargeting; bots; transparency; EU law; democratic discourse.*

INTRODUCTION[1]

The negative consequences of establishing a public forum in the online environment are now a well-established fact. It is no coincidence that already in 2018, the European Commission comprehensively addressed this phenomenon in its Communication *'Tackling Online Disinformation: A European Approach.*'[2] As the Habermasian public sphere gradually shifted from the physical world and traditional media into the digital realm, this transformation enabled new techniques for influencing public discourse and shaping mass opinion on a scale that democratic societies had never encountered before.[3]

The online space–especially through social networks and digital platforms–has made it possible to strategically create and disseminate various forms of mis-, dis-, and mal-information, giving rise to what scholars and institutions now describe as "information disorder."[4] This term encompasses the harmful effects that modern patterns of information

* JUDr. Jaroslav Denemark, Department of European Law, Faculty of Law, Charles University in Prague, Prague, Czech Republic. ORCID: 0000-0002-8304-1312. "The work was supported by the grant SVV n. 260750, International and supranational regulation of autonomization and automatization of human and machine decision-making."

[1] During the preparation of this article, the author made use of the language model ChatGPT (GPT-4, OpenAI) as a stylistic and methodological tool, particularly for linguistic refinement, structural editing, and assistance in the formulation of legal argumentation. All substantive content and legal conclusions remain the sole responsibility of the author.

[2] European Commission, Communication from the Commission to the European Parliament, the Council the European Economic and Social Committee and the Committee of the Regions, Tackling online disinformation: a European Approach, Brussels, 26. 4. 2018, COM(2018) 236 final.

[3] Ibidem, p. 5.

[4] KERMER, J. E., NIJMEIJER, R. A. Identity and European Public Spheres in the Context of Social Media and Information Disorder. *Media and Communication.* 2020, Vol. 8, No. 4; DE BLASIO, E. et al. The Ongoing Transformation of the Digital Public Sphere. *Media and Communication.* 2020, Vol. 8, No. 4, p. 34.

consumption have on global society. Truth becomes relativized, individuals are overwhelmed by the volume of content, and this phenomenon–termed an "infodemic"–was particularly evident during the COVID-19 pandemic, as highlighted by the World Health Organization.[5] Social media platforms, through algorithmic amplification and reinforced echo chambers intensify polarization across entire social groups.[6]

According to the abovementioned Communication, disinformation constitutes the most significant driver of information disorder, as it is deliberately 'created, presented and disseminated for economic gain or to intentionally deceive the public, and may cause public harm.'[7] This means that disinformation are malicious content *ipso facto*.

Disinformation is produced, amplified, and disseminated through various techniques–such as so-called fake news, fabricated media reports,[8] operations by troll factories,[9] and even internet memes (e.g., those deployed by Russia during the 2016 U.S. presidential election).[10] While these are relatively visible and recognizable techniques, other methods of creating, distributing, and amplifying disinformation are far subtler–and thus more difficult to detect and counter effectively.

The growing use of artificial intelligence in disinformation campaigns allows perpetrators to employ increasingly sophisticated strategies that often evade detection altogether, as they are designed to target the subconscious mind of the recipient rather than their conscious reasoning. The danger lies in the exploitation of psychological vulnerabilities through subtle cues, emotional triggers, or even imperceptible stimuli. Such techniques blur the boundary between persuasion and manipulation, posing novel regulatory challenges.

This article critically assesses whether the EU Artificial Intelligence Act sufficiently addresses this phenomenon of AI-enabled subliminal manipulation–or, more specifically, subliminal disinformation techniques–both in terms of conceptual clarity and regulatory enforceability.

I. MAIN SUBLIMINAL TECHNIQUES USED FOR DISINFORMATION

AI-enabled subliminal techniques are employed throughout the entire life cycle of disinformation. Artificial intelligence has made it easier than ever to create disin-

[5] See e.g. World Health Organisation, Impact of the COVID-19 infodemic on frontline workers and health systems: analysis of story-telling approach for infodemic management, 2024.

[6] CINELLI, M. et al. *"Echo chambers on social media: A comparative analysis"*. Proceedings of the National Academy of Sciences.

[7] European Commission, Communication from the Commission to the European Parliament, the Council the European Economic and Social Committee and the Committee of the Regions, Brussels, 26. 4. 2018, COM(2018) 236 final, pp. 3–4.

[8] European Commission, *Minutes of the First Meeting of the High-Level Expert Group on Fake News*. January 2018. In: *European Commission* [online]. [2025-10-07]. Available at: <https://ec.europa.eu/information_society/newsroom/image/document/2018-6/minutes_15_january_2018_meeting_hlg_fake_news_59BB-9FE9-A0B0-15BD-CB6B78C7B820F93E_49696.pdf>.

[9] European Commission, Communication from the Commission to the European Parliament, the Council the European Economic and Social Committee and the Committee of the Regions, Brussels, 26. 4. 2018, COM(2018) 236 final, p. 5.

[10] HELMUS, TODD C. Artificial Intelligence, Deepfakes, and Disinformation: A Primer. Santa Monica, p. 1. In: *RAND* [online]. 6. 7. 2022 [2025-10-07]. Available at: <https://www.rand.org/pubs/perspectives/PEA1043-1.html>.

formation.[11] Not only can AI generate social media posts, fake news articles, and other textual content, but it can also alter existing images or videos–or even generate entirely new, realistic visuals and voice recordings.

To disseminate disinformation in a targeted and effective manner, complex profiling based on behavioural psychology allows perpetrators to identify and reach those individuals who are most likely to be influenced

Finally, to amplify the reach and apparent popularity of disinformation, bots are commonly used to artificially engage with content on social media, thereby creating the illusion that it is being widely interacted with by real users.

I.1 Step one – to create

Artificial intelligence, particularly systems based on machine learning, can generate new data derived from existing images, videos, audio recordings, and other sources.[12] The creation of such content, which aims to simulate reality despite being untrue, is referred to as a deepfake.

This understanding aligns with the definition of a deepfake provided in EU law. Pursuant to Recital 134 of the AI Act, a deepfake is defined as 'image, audio or video content that appreciably resembles existing persons, objects, places, entities or events and would falsely appear to a person to be authentic or truthful.'[13]

Although the Digital Services Act does not use the term "deepfake" explicitly, it operates with a nearly identical definition. According to Article 35(1)(k), very large online platforms and search engines are required to implement reasonable, proportionate, and effective mitigation measures to ensure that an item of information, whether it constitutes a generated or manipulated image, audio or video that appreciably resembles existing persons, objects, places or other entities or events and falsely appears to a person to be authentic or truthful is distinguishable through prominent markings when presented on their online interfaces, and, in addition, providing an easy to use functionality which enables recipients of the service to indicate such information.[14]

Interestingly, the European Commission's guidelines for providers of very large online platforms and search engines concerning these mitigation measures adopt an even broader definition. They include not only fabricated content but also content that misrepresents existing persons, objects, places, entities, or events through synthetic or manipulated means.[15]

Moreover, in the Commission Recommendation on inclusive and resilient electoral processes in the Union and enhancing the European nature and efficient conduct of the

[11] European Commission, Communication from the Commission to the European Parliament, the Council the European Economic and Social Committee and the Committee of the Regions, Brussels, 26. 4. 2018, COM(2018) 236 final, p. 5.

[12] BONTRIDDER, N., POULLET, Y. *The role of artificial intelligence in disinformation. Data & Policy.* Cambridge: Cambridge University Press, 2021, pp. 32–3.

[13] AI Act, recital 134.

[14] Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market for Digital Services and amending Directive 2000/31/EC (Digital Services Act), Article 35(1)(k).

[15] Communication from the Commission – Commission Guidelines for providers of Very Large Online Platforms and Very Large Online Search Engines on the mitigation of systemic risks for electoral processes pursuant to Article 35(3) of Regulation (EU) 2022/2065, C/2024/3014, 2024, point 40.

elections to the European Parliament, deepfakes are directly linked to negative consequences for the quality of democratic debate and the integrity of electoral processes and the definition aligns with the abovementioned.[16]

A striking example of the large-scale use of deepfakes and other AI-generated disinformation occurred during the Israeli–Iranian escalation in June 2025. As reported by BBC, dozens of manipulated videos circulated across major platforms–including AI-generated images and fabricated footage falsely depicting successful Iranian strikes on Israeli targets. One such video, allegedly showing an Israeli F-35 jet being shot down, was later identified as footage from a flight simulator game.[17]

I.2 Step two – to disseminate

To maximise the effectiveness of disinformation, it must be carefully targeted and tailored to the right audience. This level of personalisation goes far beyond traditional demographic profiling, which has been a staple of political campaigning for decades. Members of the same demographic group–however narrowly defined–can still differ significantly in terms of their personality traits, opinions, values, and emotional sensitivities.[18]

Based on behavioural profiling or psychometrics, vast amounts of personal data are analysed in order to create and deliver content that is more likely to influence individuals' opinions. This psychometric targeting enables algorithms to prioritise content using an attention-based model, aiming to trigger a reaction from the user–whether through emotional engagement, perceived credibility, or active interaction such as sharing or commenting. These reactions in turn increase the content's visibility and virality within digital networks.[19]

This psychographic microtargeting is thus not only used to direct content toward selected audiences, but also to create the content itself, based on individual psychological characteristics. Such behavioural nudging can, in consequence, amount to subliminal manipulation, as the disinformation is not always immediately apparent to the recipient.

One of the most prolific and controversial cases of psychographic microtargeting in a political context is linked to Cambridge Analytica and its involvement in the 2016 U.S. presidential election campaign for Donald Trump, as well as the "Leave" campaign during the Brexit referendum. According to several whistleblowers, during the U.S. election Cambridge Analytica analysed vast quantities of online data on over 200 million U.S. citizens, including personal data from social media platforms such as Facebook, in order to determine what specific individuals liked, disliked, what their

[16] Commission Recommendation (EU) 2023/2829 of 12 December 2023 on inclusive and resilient electoral processes in the Union and enhancing the European nature and efficient conduct of the elections to the European Parliament, recital 39.

[17] MURPHY, M., ROBINSON, O., SARDARRIZADEH, S. Israel-Iran conflict unleashes wave of AI disinformation. In: *BBC* [online]. 21. 6. 2025 [2025-06-29]. Available at: <https://www.bbc.com/news/articles/c0k78715enxo>.

[18] KERTYSOVA, K. Artificial Intelligence and Disinformation: How AI Changes the Way Disinformation is Produced, Disseminated, and Can Be Countered. *Security and Human Rights*. 2018, Vol. 29, p. 64.

[19] BULKA, T. The Constitutional Implications of Regulating Microtargeting. *Fordham Intellectual Property, Media & Entertainment Law Journal*. 2022, Vol. 32, No. 4, pp. 1112–1114.

interests were, and how to overwhelm them with automatically generated, emotionally tailored content.[20]

The dangers of psychographic microtargeting have been explicitly acknowledged by the European Union since at least 2020. In its Communication on the European Democracy Action Plan, the Commission warned that such techniques 'make it much harder to hold politicians to account for the messaging and open new ways for attempts to manipulate the electorate'.[21] The use of psychographic microtargeting was clearly linked to the misuse of improperly obtained personal data to 'direct divisive and polarising narratives'.[22]

I.3 Step three – to amplify

Another powerful tool commonly employed in disinformation campaigns is the use of automated accounts, or bots. These are software-controlled profiles that can operate on social media platforms by posting, sharing, liking, and commenting at a scale and speed far beyond human capacity. Bots are primarily used to amplify content–making it appear more popular, more widely accepted, or more emotionally engaging than it actually is.[23]

Bots can range from very simple programs designed merely to like or share content, to sophisticated AI-driven systems capable of acting autonomously, commenting, interacting with users, or even generating content themselves. These more advanced bots are often indistinguishable from human users, making them especially effective at spreading disinformation covertly and at scale.[24]

By coordinating bot activity, perpetrators can simulate organic public interest, manufacture the illusion of consensus, and manipulate platform algorithms to promote disinformation more broadly. This artificial popularity may prompt real users to engage with the content, further boosting its virality and reinforcing misleading narratives. Bots can also target specific individuals with repeated exposure to the same messages, increasing their persuasive power through sheer repetition.[25]

The use of bots has been extensively exploited by Russian-led, anti-democratic, pro-Kremlin disinformation campaigns, often in connection with the political promotion of extremist or anti-establishment candidates in nationwide elections. One of the most recent examples was documented during the 2024 Croatian presidential election, in which the incumbent Zoran Milanović–a candidate known for his anti-NATO and pro-Russian stance–was re-elected.

According to independent monitoring reports, bot activity played a significant role in amplifying Milanović's campaign messaging. A coordinated disinformation effort was

[20] JAKEE, K., FINK, D. Microtargeting Voters in the 2016 US Election: Was Cambridge Analytica Really Different? In: *SSRN* [online]. 1. 5. 2024 [2025-06-30]. Available at: <https://ssrn.com/abstract=4843786>.
[21] Communication from the Commission to the European Parliament, the Council, the European Economic and Social Committee and the Committee of the Regions On the European democracy action plan, Brussels, 3. 12. 2020, COM(2020) 790 final, p. 4.
[22] Ibidem.
[23] BONTRIDDER, N., POULLET, Y. *The role of artificial intelligence in disinformation. Data & Policy,* pp. 32–5.
[24] Ibidem.
[25] Ibidem, pp. e32-7.

observed across platforms such as X (formerly Twitter) and Facebook, where Russian-affiliated bot networks supported Milanović by circulating strong pro-Kremlin narratives. One fake account identified as a bot was reportedly able to post over 100 times per day, primarily sharing visual content that either attacked opposition candidate Primorac or promoted Milanović's political positions.[26]

Furthermore, during the 2024 Romanian presidential campaign, pro-Russian candidate Călin Georgescu relied almost exclusively on TikTok as the primary platform for his outreach. Many analysts expressed surprise at how rapidly his content gained visibility and how significantly his online reach expanded.

Subsequent investigations revealed that a major factor behind Georgescu's apparent online success was the use of a sophisticated network of TikTok bots, which artificially inflated engagement metrics such as views, likes, and comments. This manipulation created the false impression that Georgescu enjoyed widespread public support, thereby influencing public perception and subliminally persuading voters that he was a viable and popular choice.[27]

Bots were already recognised as a threat by the European Union in 2018, when the European Commission's Communication 'Tackling Online Disinformation: A European Approach' identified so-called "automated services"–i.e., bots–as one of the main technology-enabled mechanisms used to artificially amplify the spread of disinformation. The Communication highlighted that such services can be used to boost the visibility of false or misleading content, create the illusion of popular support, and manipulate both public discourse and online platform algorithms.[28]

II. RESPONSE OF ARTIFICIAL INTELLIGENCE ACT

The Artificial Intelligence Act adopted in 2024, represents the European Union's first horizontal regulatory framework for artificial intelligence. While its primary aim is to ensure the safe and trustworthy development of AI systems, the regulations also touches upon the use of AI for the purpose of creating, disseminating, and amplifying disinformation.

The risks posed by AI in this context are explicitly acknowledged in several recitals of the Act. It is stated that general-purpose AI models may pose systemic risks, including, among others, the generation and spread of false or misleading content and the facilitation of disinformation campaigns.[29]

---

[26] STARCEVIC, S. Russian bots boosted NATO critic ahead of Croatian election, researches say. In: *Politico* [online]. [2025-06-30]. Available at: <https://www.politico.eu/article/russia-bots-nato-croatia-election-presidential-candidate-eu-donald-trump-zoran-milanovic-campaign/#:~:text=Russian%20bots%20launched%20a%20%E2%80%9Cpro,group%20of%20researchers%20said%20Wednesday>.

[27] MIHAILESCHU, D. How Romania's Presidential Election Became the Plot of a Cyber-Thriller. In: *European Union* [online]. [2025-06-30]. Available at: <https://youth.europa.eu/news/how-romanias-presidential-election-became-plot-of-cyber-thriller_en#:~:text=The%20key%20tool%3F%20TikTok%20bots%2C,Court%20did%20not%20cancel%20the>.

[28] European Commission, Communication from the Commission to the European Parliament, the Council the European Economic and Social Committee and the Committee of the Regions, Tackling online disinformation: a European Approach, Brussels, 26. 4. 2018, COM(2018) 236 final, p. 5.

[29] AI Act, recital 110.

Furthermore, Recitals 120 and 136 highlight that providers of very large online platforms and search engines–in line with their obligations under the Digital Services Act–are required to identify and mitigate systemic risks arising from the dissemination and amplification of AI-generated disinformation. The AI Act's recitals also specify transparency obligations with respect to synthetic content, especially deepfakes, to ensure that users can distinguish such content from authentic material.[30]

Subliminal techniques deployed by AI systems are addressed directly in Article 5(1)(a) of the AI Act, which classifies them as a prohibited practice. According to this provision, it is prohibited to place on the market, put into service, or use 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.*[31]

This provision aims to protect individuals and groups from non-consensual manipulation that impairs their autonomy and decision-making capacity–even in the absence of intent, where the effect alone is sufficient to trigger the prohibition. The reference to "significant harm" plays a crucial role in interpreting the scope of this article and will be further discussed below.

This prerogative is further elaborated in Recital 29 of the AI Act, which also offers a definition of subliminal techniques as stimuli that 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 not able to control or resist them.[32]

Subliminal techniques are particularly dangerous because they can prompt individuals to behave in ways they otherwise would not or subtly nudge them towards certain beliefs or behaviours–thereby undermining their free will and the right to make autonomous choices.

Recital 29 also implicitly refers to psychographic microtargeting, especially when it stresses that subliminal techniques are most effective when directed at vulnerable individuals. It highlights specific vulnerabilities such as extreme poverty, minority status (ethnic or religious), disability, or the fact that a person is a minor, which can all amplify the impact of manipulative AI systems and increase the risk of exploitation.

Needless to say, subliminal practices are not prohibited per se. According to Article 5(1)(a), they are prohibited only where their use causes or is reasonably likely to cause significant harm to the affected person, another person, or a group of persons.[33] The concept of "harm" is not expressly defined in Article 5 itself. However, Recital 29 sug-

[30] Ibidem, recital 134.
[31] Ibidem, Article 5(1)(a).
[32] Ibidem, recital 29.
[33] Ibidem, Article 5(1)(a).

gests a broad interpretation by including examples such as economic or financial harm. This indicates that the notion of harm should not be limited solely to physical injury or immediate psychological effects but may extend to less tangible yet equally serious interferences, particularly where they affect a person's free will, autonomy, or informed decision-making.

Nonetheless, such an extensive interpretation of "harm" remains legally uncertain, as it has not yet been explicitly confirmed by the Court of Justice of the European Union (CJEU). This ambiguity is compounded by the fact that Article 3(6) of the AI Act provides a more specified framed definition of harm in the context of high-risk systems–namely, as 'harm to the health, safety or fundamental rights of natural persons.'[34]

Article 6 of the AI Act defines the scope of high-risk AI systems, primarily by reference to the sectors and use cases listed in Annex III. Subliminal techniques as such do not fall within the categories defined in Article 6(1), nor are they currently included in Annex III. In other words, AI systems that deploy subliminal techniques for the purpose of creating, disseminating, or amplifying disinformation are not classified as high-risk AI systems, and thus the regulatory framework applicable to high-risk systems does not currently apply to them.

However, Article 7(1)(b) provides that the European Commission is empowered to adopt delegated acts to amend Annex III. This allows the Commission to add new use cases if an AI system poses a risk to health and safety or an adverse impact on fundamental rights, and such risk is equivalent to or greater than that posed by the existing high-risk categories already listed. This mechanism offers a potential pathway for future inclusion of AI-driven disinformation tools–particularly those relying on subliminal techniques–within the high-risk category, should their impact on fundamental rights be sufficiently recognised.[35]

Article 50 of the AI Act establishes transparency obligations for both providers and deployers of certain AI systems. Under the Act, a provider is defined as an entity that develops an AI system and places it on the market or puts it into service under its own name or trademark, while a deployer is an entity that uses an AI system under its authority.[36]

These definitions imply that providers of very large online platforms (VLOPs) and search engines–as regulated primarily under the Digital Services Act–do not qualify as "providers" or "deployers" under the AI Act in relation to content made publicly available solely via their services, unless they are directly responsible for the development or use of the AI system in question. As a result, AI-generated content disseminated exclusively through such platforms is not automatically subject to the transparency requirements of Article 50, unless the platform itself acts as the provider or deployer of the AI system used to generate that content.

However, pursuant to Article 50(1) of the AI Act, providers must ensure that natural persons are informed that they are interacting with an AI system, unless this is obvious from the context and the person's circumstances. This obligation could, in principle, be applicable to AI-powered bots simulating human interaction.

[34] Ibidem, Article 6(3).
[35] Ibidem, Article 7(1)(b).
[36] Ibidem, Article 3(3), (4).

Nevertheless, very large online platforms (VLOPs)–which often serve as hosts for such bots–are not considered providers or deployers under the AI Act, unless they develop or actively use the AI systems themselves. Consequently, they are not directly obliged under Article 50 to monitor, detect, or label AI-generated interactions or fake accounts operated by third parties. This creates a potential regulatory gap, particularly with regard to inauthentic behaviour and covert use of AI in disinformation campaigns, which may go undetected on large platforms unless addressed under other legal frameworks, such as the Digital Services Act.

Article 50(2) of the AI Act requires providers to ensure that outputs generated by AI systems are marked in a machine-readable format and are detectable as artificially generated or manipulated. In other words, AI-generated content must be labelled, but the obligation applies only to machine-readability–there is no requirement that such labelling be visible or understandable to human users.

By contrast, Article 50(4) extends the transparency obligation to deployers in the specific context of deepfakes. In such cases, deployers are required to disclose that the content was artificially generated or manipulated in a manner that is perceptible to natural persons, not merely through machine-readable metadata. This aims to ensure that human users are not misled by synthetic or manipulated media.

However, the obligation under Article 50(4) is subject to a significant exception: it does not apply where the content is evidently artistic, creative, satirical, fictional, or analogous in nature. This open-ended exemption may significantly limit the effectiveness of the provision, as disinformative or manipulative content could easily be presented under the guise of satire, fiction, or artistic expression–thereby circumventing the duty to disclose its synthetic origin.

Pursuant to Article 51 of the AI Act, certain general-purpose AI models may be classified as general-purpose AI models with systemic risk. This classification applies where the model, due to its technical capabilities and widespread influence, poses a significant risk to the internal market, public safety, or fundamental rights. In particular, such classification may arise when the model is used to generate or manipulate content for the purpose of subliminal techniques, especially where this use has a high impact on individuals or society.

Recital 111 provides further guidance by identifying several criteria for such assessment, including the model's reach, the cumulative amount of computation used for training, algorithmic advancements, hardware efficiency, model capability, number of business and end users, and level of autonomy. These factors serve as a basis for evaluating whether a general-purpose AI model should be designated as posing systemic risk–either through independent technical assessment or by a formal decision of the European Commission.

As of today, the most likely candidates for such classification include GPT-4 (as used in ChatGPT), Gemini Ultra, and Claude 3 Opus, given their scale, technical sophistication, and broad integration across commercial and public domains.

Providers of general-purpose AI models with systemic risk are, under the AI Act, subject to a limited and broadly formulated set of obligations, such as conducting model evaluations, and assessing and mitigating systemic risks associated with the use of their models (Articles 53–55). However, these obligations are vague in scope and lack specific

guidance, particularly when it comes to the deployment of such models for subliminal techniques.

As a result, it remains unclear whether and to what extent the use of systemic-risk general-purpose AI models in the creation or dissemination of subliminal disinformation is adequately covered by the current regulatory requirements. The absence of concrete benchmarks or evaluative criteria leaves a considerable gap in enforceability and legal certainty.

Moreover, no other provisions of the AI Act expressly address the use of AI for subliminal disinformation techniques, unless such systems are explicitly designed for that purpose and fall under Article 5(1)(a) (prohibited practices). This regulatory silence underscores a potential loophole in the AI Act's ability to comprehensively address covert manipulative uses of AI in the disinformation context.

Although this chapter focuses solely on the AI Act, it should be noted that certain regulatory shortcomings may be partially mitigated by other instruments of EU law–notably the Digital Services Act (DSA), which imposes systemic risk mitigation duties on very large online platforms, and the General Data Protection Regulation (GDPR), which governs the processing of personal data including profiling and behavioural targeting. A comprehensive analysis of these regimes lies outside the scope of this chapter but remains essential for a full understanding of the EU's legal approach to AI-enabled disinformation.

CONCLUSION

This article analysed the regulation of AI-enabled subliminal techniques used in the creation, amplification, and dissemination of disinformation under the AI Act. The analysis focused exclusively on the AI Act's legal framework, without drawing on the GDPR or the Digital Services Act, although these instruments may help address some of the regulatory gaps identified.

The research was carried out through a doctrinal analysis of the AI Act's relevant provisions and their supporting recitals. Particular attention was paid to the definition and scope of "subliminal techniques," the prohibition under Article 5(1)(a), and the concept of "harm." The findings suggest that the AI Act acknowledges the risks posed by AI-generated disinformation but only partially addresses them. Subliminal manipulation is prohibited under Article 5(1)(a), but only when it causes or is likely to cause "significant harm." This harm is not clearly defined, and its interpretation is complicated by the more limited definition of harm in Article 3(6). Disinformation techniques based on subliminal AI manipulation do not fall under the current high-risk classification, and the vague obligations imposed on systemic-risk general-purpose AI models offer little legal certainty.

Transparency rules under Article 50 are weakened by narrow personal and material scope, as well as broad exemptions–particularly for deepfakes that may be labelled as satirical or artistic. Moreover, very large online platforms and search engines are not considered providers or deployers under the AI Act unless they themselves develop or use the AI tools, creating further enforcement gaps.

While the AI Act offers important tools to address certain forms of manipulation, it falls short of comprehensively regulating the use of subliminal AI techniques in the disinformation context. This is a serious blind spot given the growing sophistication of AI-enabled persuasion techniques and their documented use in recent political campaigns.