# Disinformazione digitale

## Modelli algoritmici e valori democratici

a cura di
Licia Califano
Federica Fabrizzi
Giovanni Sartor

Collana di Diritto

FrancoAngeli

SAGGI E RICERCHE

 ISBN 9788835191209




Copyright © 2026 by FrancoAngeli s.r.l., Milano, Italy. ISBN 9788835191209

# Disinformazione digitale

## Modelli algoritmici e valori democratici

a cura di
Licia Califano
Federica Fabrizzi
Giovanni Sartor

FrancoAngeli

Collana di Diritto

SAGGI E RICERCHE

 ISBN 9788835191209


Questo volume rappresenta il risultato conclusivo del Progetto PRIN 2022 PNRR "DAFNE" (*Democratic governance of Automated system for Fake News*), finanziato dall'Unione europea – Next Generation EU nell'ambito del PNRR, Missione 4 Componente 2, Investimento 1.1, Bando "PRIN 2022 PNRR" del MUR emanato con Decreto Direttoriale n. 1409 del 14 settembre 2022, Codice MUR P2022R7RS9, CUP H53D23010930001, e raccoglie alcune delle relazioni e degli interventi presentati al Convegno "Poteri pubblici e poteri privati nel governo dell'informazione digitale: la scelta europea dell'accountability" tenutosi presso l'Università degli Studi di Urbino Carlo Bo il 24-25 ottobre 2025.

Isbn e-book Open Access: 9788835191209



Copyright © 2026 by FrancoAngeli s.r.l., Milano, Italy. ISBN 9788835191209

# Indice

**Sezione II**
**Informazione digitale**
**e tutela dell'individuo**



**Sezione III**
**Comunicazione elettorale e politica**





 ISBN 9788835191209

**Riflessioni conclusive**

# The Governance of "Augmented Disinformation" between the Digital Services Act and the AI Act

*Federico Galli*, Giuseppe Contissa***



## 1. Misinformation and the problem with AI

The integration of AI into digital communication has revolutionised the way information is produced, distributed and received. This transformation has also brought considerable risks, especially in terms of the amplification and dissemination of false or manipulative content.

Today, AI plays a crucial role in amplifying disinformation[1] through the use

* Ricercatore di Filosofia del diritto, Alma Mater Studiorum – Università di Bologna.

** Professore associato di Filosofia del diritto, Alma Mater Studiorum – Università di Bologna.

1. We will not enter the debate of defining "disinformation" as well as of the inadequacy of the concept of "fake news". See O. Pollicino, *Delegation of Regulatory Functions and Combating Systemic Risks Caused by Disinformation in the Digital Services Act: What Risks to Freedom of Expression?*, in *MediaLaws*, n. 3, 2023, pp. 114-43. As many now do, we prefer to approach the issue using the definition provided by the High-Level Expert Group, namely «false, inaccurate or misleading information, designed, presented and promoted with the intent to cause public harm or generate profit». High Level Expert Group on Fake News and Online Disinformation, A *Multi-Dimensional Approach to Disinformation: Report of the Independent High-Level Group on Fake News and Online Disinformation*, European Commission, 2018.

of recommendation algorithms, which structure the visibility of content based on engagement and emotional response metrics[2]. Even when such algorithms are not designed with the intent to spread falsehoods, their operation incentivises the visibility of controversial, polarising or sensationalist content that may coincide with disinformation.

According to the University of Oxford's Computational Propaganda Research Project report, many platforms use algorithms that can unknowingly promote the spread of disinformation by rewarding content that generates more interactions, regardless of its accuracy[3]. Another study, published in *Science*, provided empirical evidence that fake news spreads faster and more widely than real news, precisely because of its emotional charge and novelty, which are two dimensions that AI algorithms leverage[4].

Another critical vector of disinformation is the use of AI for user profiling and *micro-targeting* of disinformation content[5]. Digital platforms use AI and machine learning techniques to collect behavioural, demographic and psychometric data to build users' profiles, which are then used to target tailored content. This capability can be exploited for the benefit of end users (e.g., when positive and relevant content is targeted) or it can also be exploited for malicious purposes. In this case, for example, disinformation content can be "personalised" based on the biases, fears and prejudices of individuals, thus making it more credible and persuasive. The most well-known case is certainly that of Cambridge Analytica, where data collected via Facebook was used to influence political opinions through targeted campaigns[6].

2. See, for example, V. Bakir, A. McStay, *Empathic Media, Emotional AI, and the Optimisation of Disinformation*, in M. Boler, E. Davis (eds.), *Affective Politics of Digital Media*, Routledge, 2020, pp. 263-79, where the authors analyse how "empathic media", driven by AI, allow biometric and behavioural signals – such as facial expressions, tone of voice, typing rhythm or social media interactions – to be exploited to spread disinformation.

3. S. Bradshaw, P.N. Howard, *The Global Disinformation Order: 2019 Global Inventory of Organised Social Media Manipulation*, Oxford Internet Institute, University of Oxford, 2019, p. 20.

4. S. Vosoughi, D. Roy, S. Aral, *The Spread of True and False News Online*, in *Science*, n. 6380, 2018, pp. 1146-1151.

5. F. Lagioia, G. Sartor, *Profilazione e Decisione Algoritmica: Dal Mercato Alla Sfera Pubblica*, in *federalismi.it*, n. 11, 2020, pp. 85-110.

6. Among the various contributions that have analysed this case, see the interview conducted by Carole Cadwalladr with Christopher Wylie, the Cambridge Analytica whistleblower, who revealed how data from over 50 million Facebook profiles was collected without consent to feed a psychometric profiling system aimed at large-scale electoral manipulation. The investigation, published by The Guardian, highlighted the central role of the misuse of data and AI-assisted behavioural targeting techniques in influencing the political opinions of American voters during the 2016 elections, and represented a turning point in public awareness of the systemic risks associated with algorithmic surveillance and personalised disinformation. C. Cadwalladr,

In electoral or healthcare contexts, personalised disinformation can have a direct effect on individual choices, fuelling abstentionism, radicalisation or distrust of institutions. This is a delicate area, where AI does not simply spread disinformation, but also adapts it to the individual user, reinforcing *echo chambers* and further polarising public debate[7].

AI techniques are now also used to create advanced bots, i.e. programmes capable of interacting with other online users by simulating human behaviour. These bots, which can operate on a large scale, are used to spread disinformation, manipulate trends, create fake "engagement" or fuel coordinated campaigns.

For example, it has been shown that bots played a central role in spreading disinformation during political events such as the 2016 American elections or the Brexit referendum[8]. Bots can also amplify the effect of "astroturfing", i.e. simulating spontaneous grassroots consensus that is actually artificial and planned. In addition to spreading disinformation, bots can also play an active role in attacking legitimate sources, delegitimising journalists, scientists or activists through automated harassment campaigns, discrediting them or spreading conspiracy theories.

Disinformation is therefore not only conveyed through AI, but is directly produced by AI systems, contributing to the creation of an "information epidemic"[9] or "information disorder"[10] where media noise outweighs the merit of communication.

Finally, in the field of fake content production, the recent spread of generative AI models, such as Gpt, Claude or Gemini, in large-scale applications, such as ChatGpt, has taken the relationship between AI and disinformation to a new level. These systems, which enable the autonomous production of textual, visual and audiovisual content, are capable of creating fake news, manipulated narratives, realistic but non-existent images, and even videos in which public figures appear to say or do things that never happened (so-called *deepfakes*)[11].

E. Graham-Harrison, *Revealed: 50 Million Facebook Profiles Harvested for Cambridge Analytica in Major Data Breach*, in *The Guardian*, 2018, *www.theguardian.com/news/2018/mar/17/cambridge-analytica-facebook-influence-us-election*.

7. C.R. Sunstein, *#Republic: Divided Democracy in the Age of Social Media,* Princeton University Press, 2017, p. 20.

8. E. Ferrara *et al*., *The Rise of Social Bots,* in *Communications of the ACM*, n. 7, 2016, pp. 96-104.

9. C. Wardle, H. Derakhshan, *Information Disorder: Toward an Interdisciplinary Framework for Research and Policymaking*, Council of Europe, 2017.

10. R. Bracciale, F. Grisolia, *Information Disorder: Acceleratori Tecnologici e Dinamiche Sociali*, in *federalismi.it*, n. 11, 2020, pp. 58-72.

11. A.M. Walorska, *Deepfakes and Disinformation*, Friedrich Naumann Foundation for Freedom, 2020.

According to Chesney and Citron, *deepfakes* represent an "epistemic threat"[12], as they undermine the very concept of "shared reality": if everything can be realistically faked, it becomes difficult to believe even the truth. This effect is called the "*liar's* dividend", i.e. the possibility for anyone to deny documented facts by claiming that they are manipulated[13].

At the same time, *large language models* (Llm) can be used to produce large-scale disinformation text content consistently and in a matter of seconds. As Luciano Floridi observes, the ability of these models to simulate human language without understanding the content generates a structural risk: the texts produced may be grammatically correct and stylistically persuasive, but the semantics of the messages and their veracity are completely detached from reality[14]. When used maliciously, they therefore become powerful tools for the large-scale production of propaganda or fake news.

## 2. The solution with AI?

Although artificial intelligence is in the spotlight for its role in spreading and creating disinformation, it is also a tool with extraordinary potential to counteract its negative effects. Today, most initiatives aimed at combating the spread of disinformation take place thanks to AI technologies.

One of the main contributions of AI to the fight against disinformation is its ability to automatically detect false or misleading content. Thanks to Natural Language Processing (Nlp) models, AI can analyse huge amounts of textual data to identify linguistic signals characteristic of disinformation: excessive generalisations, sensationalist statements, alarmist tones or arguments without verifiable sources.

These systems mainly rely on supervised classification models, which learn to distinguish between true and false news from large manually labelled datasets[15]. During training, a learning algorithm receives examples of

12. R. Chesney, D. Citron, *Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security*, in *California Law Review*, n. 6, 2019, pp. 1753-1819.

13. K.J. Schiff, D.S. Schiff, N.S. Bueno, *The Liar's Dividend: Can Politicians Claim Misinformation to Evade Accountability?*, in *American Political Science Review*, n. 1, 2025, pp. 71-90. The study highlights how this practice can compromise politicians' accountability. In particular, the authors identify two main strategies: invoking informational uncertainty and encouraging opposition support from their supporters.

14. L. Floridi, *The False Promise of ChatGpt*, in *Philosophy & Technology*, n. 3, 2023, pp. 1-6.

15. For an up-to-date overview of *fake news* analysis models, see X. Zhou, R. Zafarani, *A Survey of Fake News: Fundamental Theories, Detection Methods, and Opportunities*, in *ACM Computing Surveys*, n. 5, 2020, pp. 1-40.

previously verified content (labelled as "true" or "false") and learns to identify correlations between textual characteristics (such as syntactic structure, the presence of hyperbolic terms, the absence of references to reliable sources, or the repeated use of polarising rhetoric) and the veracity of the content. Once trained, the model can be used in the operational phase to analyse new texts and estimate, with a certain degree of probability, whether they are potentially disinformation.

In addition, AI can compare the content of a message with databases of verified news or official statements to identify its plausibility. Several studies have confirmed the effectiveness of these approaches. For example, it has been shown that machine learning can be used to recognise fake news on social media with a good level of accuracy[16]. This preventive screening function, especially when integrated into platform moderation systems, is a powerful tool for limiting the viral spread of disinformation.

In this field, AI often plays a supporting role to professionals in the fact-checking industry, speeding up manual verification. AI systems can automatically retrieve official sources, previous articles or public data to compare with suspicious claims, significantly reducing the human workload. Some tools, such as ClaimBuster or Full Fact AI, already used by news outlets or in academic settings, allow verifiable statements to be identified in real time during political events or interviews.

Another area where AI has proven particularly effective is in mapping disinformation networks, i.e. identifying accounts, pages and groups that act in a coordinated manner to spread false or manipulative content. Through social network analysis and clustering techniques, algorithms can analyse content sharing patterns, publication timings and message repetitiveness to uncover suspicious connections between seemingly independent subjects[17].

Finally, recent developments in generative AI have also found application in the field of disinformation control. For example, Large Language Models (Llm) have been progressively used as tools to verify suspicious claims, synthesise information from authoritative sources, and provide clear and contextualised explanations on controversial issues. Some of these models are capable of

16. K. Shu *et al.*, *Fake News Detection on Social Media: A Data Mining Perspective*, in *ACM SIGKDD Explorations Newsletter*, n. 1, 2017, pp. 22-36. The study highlights how the combined analysis of textual content, social data (such as interactions and retweets) and metadata related to sources can significantly improve the ability of predictive models to distinguish real news from fake news, paving the way for the integration of these tools into digital platform monitoring systems.

17. T. Papanastasiou, *Fake News Propagation and Detection: A Case Study,* in *Online Social Networks and Media*, n. 20, 2020, p. 100103.

processing a request in natural language, comparing it with reliable sources and producing a response that indicates whether or not a given statement is reliable.

An interesting case in point is Grok, the Llm-based chatbot integrated into the X platform (formerly Twitter), developed by xAI. Starting in 2024, several users began using Grok as a sort of on-demand fact-checker, querying it directly under controversial posts to obtain immediate verification of the content[18]. This practice has raised significant debates: on the one hand, it demonstrates growing confidence in the potential of generative AI as a tool against misinformation; on the other, it raises concerns about its reliability, transparency and possible spread of errors.

## 3. The risks of "augmented disinformation"

By "augmented disinformation" we mean the set of transformations that the phenomenon of disinformation undergoes as a result of the integration of AI into the processes of production, dissemination and management of information content. AI acts as a powerful multiplier, capable of making disinformation faster, more personalised and more convincing, while also offering effective tools to combat it. In this sense, it represents a true "Janus face" of the contemporary (dis)information system: on the one hand, it enables new threats, while on the other, it supports the defence of truthful information.

It is precisely this ambivalence that calls for careful reflection on how to govern the use of AI in combating disinformation, since the technologies used to protect the information ecosystem can, in turn, generate unintended, sometimes serious, distortive effects on the quality of information itself, fundamental rights, pluralism and the democratic quality of public debate.

AI systems used to combat disinformation are mainly based on machine learning techniques and natural language models, designed to identify recurring patterns in content and classify them according to their reliability. However, these systems do not operate through a deep semantic understanding of content, but through statistical correlations and probabilistic models trained on large volumes of data. This implies a number of significant risks ranging from technical errors to broader systemic effects.

18. J. Singh, *X Users Treating Grok Like a Fact-Checker Spark Concerns over Misinformation*, in *TechCrunch*, 2025, *techcrunch.com/2025/03/19/x-users-treating-grok-like-a-fact-checker-spark-concerns-over-misinformation/*.

One of the main risks concerns classification errors, in particular the possibility that legitimate content may be mistakenly labelled as false or harmful. Automatic moderation systems, even if advanced in terms of semantic language analysis (e.g. through word embeddings), do not understand the semantic referents of words, let alone the intention behind the dissemination of content. This could lead, for example, to the removal of satirical, critical or simply controversial content that is perfectly lawful; or to the removal of false content that the user believes to be true in good faith (which would be a case of misinformation). This is a phenomenon known as *overblocking* or *collateral censorship*, which can seriously compromise users' freedom of expression, especially in politically sensitive areas[19].

Furthermore, AI's ability to grasp cultural, linguistic or rhetorical nuances is still limited, making it difficult to distinguish between misinformation and unconventional forms of opinion. In a digital space where the margin for error can have systemic effects, relying too heavily on automated decisions can be problematic, if not harmful.

Added to this is the well-documented risk of algorithmic bias[20]. AI tools are trained on large amounts of data, and if this data reflects pre-existing biases (cultural, linguistic, ideological) it is likely that these distortions will be replicated or even amplified by the models. In the context of disinformation, this can mean that certain social groups or linguistic communities are disproportionately penalised[21]. Content conveyed through low-resource languages, by political or religious minorities, or by individuals who use non-standard registers of communication, may be considered less reliable, if not directly disinformative. This raises questions in terms of fairness and inclusivity, as algorithmic moderation could end up further marginalising voices that are already underrepresented in public debate.

19. J. Balkin, *Free Speech in the Algorithmic Society: Big Data, Private Governance, and New School Speech Regulation,* in *University of California Davis Law Review*, n. 3, 2018, p. 1149. See also M. Durante, *Computational Power: The Impact of Ict on Law, Society and Knowledge,* Routledge, 2021, p. 124 where the Author analyses the various ways of "regulating" information by Internet intermediaries.

20. The literature on this subject is vast. Reference should be made to one of the first writings on the subject, C. O'Neil, *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy,* Crown, 2017; S. Quintarelli, *et al*., *AI: Ethical Profiles. An Ethical Perspective on Artificial Intelligence: Principles, Rights and Recommendations*, in *BioLaw Journal*, n. 3, 2019, pp. 159-177.

21. See S. Park *et al*., *AI Chatbots and Linguistic Injustice,* in *Journal of Universal Language*, n. 1, 2024, pp. 99-119, which highlights how AI-based chatbots can reinforce linguistic injustices, penalising speakers of minority or non-standard languages in interactions mediated by artificial intelligence.

A further problem is the opacity that characterises many AI systems, particularly those based on deep neural networks. These technologies are often described as "black boxes" precisely because, despite offering high performance, they do not make it easy to understand the logical path that led to a particular decision[22]. This means that if content is automatically flagged as "misinformation", a user may see their post removed or penalised without knowing exactly why and without having the means to effectively challenge the decision. This raises not only transparency issues, but also principles of procedural justice, as any control system should provide for the possibility of review and appeal.

The use of AI to select, filter or remove disinformation content can have a significant impact on information pluralism, even in the absence of malicious intent. Where artificial intelligence is trained to promote content considered more "safe" or "authoritative", there is a risk that the variety of opinions will be drastically reduced, especially in contexts where the truth is not unambiguous but subject to debate[23]. This is particularly true in areas such as political communication, science dissemination practices or economic theories where there are legitimate differences in the interpretation of facts. If algorithms tend to favour institutional or "mainstream" sources, critical content or alternative voices risk being suppressed even when they are not examples of disinformation at all. The danger, therefore, is that AI will become a tool for standardising public discourse, where complexity is sacrificed in the name of information security.

Finally, the automation of the fight against disinformation presents the risk (in today's political context, far from hypothetical) of instrumental use by private or public actors. On the one hand, online platform providers could use AI systems to moderate content in accordance with their economic or commercial interests, penalising information that hinders their political agenda or that of their investors. On the other hand, authoritarian governments (but not only) could exploit the automatic labelling of disinformation to target political opponents, activists or journalists deemed inconvenient, masking what is in reality likely to be selective censorship as "online environment security". In this sense, AI is an ambiguous technology: it can protect democratic debate, but it can also be used to manipulate it if it is not accompanied by appropriate legal safeguards, independent oversight bodies and civil society participation.

22. F. Pasquale, *The Black Box Society: The Secret Algorithms That Control Money and Information*, Harvard University Press, 2015.

23. G. De Gregorio, N. Stremlau, *Inequalities and Content Moderation,* in *Global Policy*, n. 5, 2023, pp. 870-879.

## 4. Augmented disinformation in the Digital Services Act

EU Regulation 2022/2065, known as the "Digital Services Act" (Dsa), was officially adopted on 19 October 2022 and entered into force on 16 November 2022, with full application from 17 February 2024, with the exception of certain obligations for large platforms applied from August 2023[24]. Included alongside the Digital Markets Act (Dma) in the broader European strategic plan for regulating the digital space, the Dsa aims to provide a regulatory framework to ensure a safer, more transparent online ecosystem that respects fundamental rights[25].

In public debate, the Dsa is often presented as one of, if not the, new legal frameworks created by the EU to combat online disinformation[26]. On closer inspection, however, the Dsa does not address the issue of automatic moderation of disinformation in a direct and systematic manner[27].

Firstly, the term "disinformation" never appears in the text of the Dsa. The only occurrences are contained in the recitals, which refer to the concept in combination with other semantically related terms: "online disinformation" (Recital 2), "disinformation campaigns" (Recital 69) and "coordinated disinformation campaigns" (Recital 83).

Moreover, in these cases, the uses of the term "disinformation" seem to reflect different meanings. For example, Recital 2 refers to "online disinformation" as a "social risk" and Recital 83 mentions "coordinated disinformation campaigns" as an example of "systemic risks" related to public health. Meanwhile, Recital 69 refers to "manipulative techniques" that contribute to disinformation campaigns and «can negatively impact entire groups and amplify societal harms».

In reality, as we will see in the following paragraphs, the phenomenon of disinformation, particularly in the sense discussed in the previous section, has an indirect relevance to the application of the Dsa.

24. 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), OJ L 277, 27.10.2022, pp. 1-102.

25. F. Casolari, *Il "Digital Services Act" e La Costituzionalizzazione Dello Spazio Digitale Europeo,* in *Giurisprudenza Italiana*, n. 2, 2024, pp. 462-465.

26. Z. Meyers, *Will the Digital Services Act Save Europe from Disinformation?*, Centre for European Reform, 20 April 2023, *www.cer.eu/insights/will-digital-services-act-save-europe-disinformation*; Deutsche Telekom Ag, *Digital Services Act (Dsa)*, Deutsche Telekom, 13 July 2023, *www.telekom.com/en/company/details/digital-services-act-dsa-1060552*; C. Goujard, *EU Wants Big Tech to Help Deepfake-Proof the 2024 Election*, Politico, 14 February 2024, *www.politico.eu/article/eu-big-tech-help-deepfake-proof-election-2024*.

27. M. Husovec, *The Digital Services Act's Red Line: What the Commission Can and Cannot Do about Disinformation*, in *Journal of Media Law*, n. 1, 2024, pp. 47-56.

The starting point for this line of reasoning is that the phenomenon can be examined in the Dsa, first and foremost, in terms of its material component, as it manifests itself in the publication of unlawful, or rather "illegal", content, which intermediary service providers are required to remove diligently. We will see, in fact, that this possibility may also apply to disinformation content that is not unlawful but is incompatible with the terms of use of the service.

In such cases, service providers may – and in a sense are encouraged to – lawfully initiate voluntary investigations to locate, identify and remove content, provided that this is done in compliance with the due diligence obligations set out in Chapter III of the Regulation. We will see that these specifically consider cases where "content moderation" is carried out with the aid of automated systems and, in this regard, provide certain safeguards for users.

Finally, for service providers that are Very Large Online Platforms (Vlop) or Very Large Online Search Engines (Vlose), disinformation is considered a potential systemic risk, particularly in relation to the design or operation of their service and its related systems. In such cases, Vlop and Vlose must carry out risk assessment and mitigation activities and comply with a number of specific obligations.

### *4.1. Disinformation content as "illegal content"*

Article 3(h) of the Dsa defines "illegal content" as «any information that, in itself or in relation to an activity, including the sale of products or the provision of services, is not in compliance with Union law or the law of any Member State which is in compliance with Union law, irrespective of the precise subject matter or nature of that law».

Given this definition, a key point in establishing the scope of the Dsa is to understand the extent to which the law of a Member State classifies disinformation as illegal content.

We know, for example, that there is no legislation in the Italian legal system that directly classifies disinformation as illegal content. However, there are sectoral regulations that could be relevant. A case in point is the dissemination of false news likely to disturb public order (e.g. Article 656 of the Italian Criminal Code)[28], or the legislation on misleading advertising, i.e. advertising

28. See S. Flaminio, *Lotta Alle Fake News: Dallo Stato Dell'arte a Una Prospettiva Di Regolamentazione Per Il "Vivere Digitale" a Margine Del Digital Services Act,* in *Rivista Italiana di Informatica e Diritto*, n. 2, 2022, p. 82, on other criminal offences that may involve the dissemination of false news.

that provides false information that may distort the behaviour of the average consumer (Article 20 of the Consumer Code). In such contexts, disinformation may take on legal significance and constitute unlawful content.

Another relevant example concerns recent legislation in some European countries regarding the dissemination of false information about the effects of Covid-19 vaccines. For example, in Germany, specific measures have been introduced to combat health misinformation, particularly on social media, with penalties for the dissemination of content that incites mistrust of vaccination campaigns. In countries such as Poland, on the other hand, the approach has been more cautious and less restrictive, leaving more room for freedom of expression even on controversial issues.

Phenomena such as these raise considerable difficulties in the uniform implementation of Dsa regulations, as online platforms may find themselves in a position where they have to remove content in one country while allowing it to be disseminated in another.

If misinformation is classified as illegal content under national law, then the provisions of the Dsa apply, relating, on the one hand, to the exemption from liability for intermediary service providers and, on the other, to specific obligations in combating illegal content.

With regard to liability, Articles 4, 5 and 6 apply, which reproduce, essentially unchanged from the E-Commerce Directive[29], the conditions under which intermediary service providers (in particular providers of mere conduit, temporary storage and hosting services, including online platforms) may benefit from immunity from non-contractual liability for illegal content transmitted or stored through their services. In the case of hosting service providers, for example, immunity applies provided that the provider is not aware of the illegal activity or content or, as soon as it becomes aware of the illegality, acts promptly to remove it or disable access to it.

In the case of illegal disinformation content, this means that a provider can only be exempt from liability if it acts promptly once it has received a report from a user or a notification from the competent authority. In the case of an order issued by a competent authority (whether judicial or administrative), Article 9 of the Dsa applies in particular, requiring intermediary service providers to cooperate with the authorities in combating illegal content. In particular, they must designate a contact point to receive removal orders issued by the authorities of Member States and act in accordance with such requests.

29. G. Giordano, *La responsabilità degli intermediari digitali nell'architettura del "Digital Services Act": è necessario che tutto cambi affinché tutto rimanga com'è?*, in *Comparazione e Diritto Civile*, n. 1, 2023, pp. 193-221.

In addition to the obligations to remove content following receipt of an order from the authorities, the Dsa has introduced a "notice and action" mechanism for hosting service providers, which was not included in the 2000 Directive[30]. This procedure should allow any user of the service to report content that is considered illegal. The report must contain a sufficiently detailed explanation, allowing the provider to assess the illegal nature of the content, and must be accompanied by the information necessary to accurately identify the subject of the request. If the report meets all the formal criteria, it has the legal effect of giving the provider actual knowledge of the unlawful content, within the meaning of Article 6 of the Dsa, thereby triggering its obligation to take prompt action.

The effectiveness of the reporting system is reinforced by Article 22, which introduces the concept of "trusted flaggers". These generally are bodies with proven experience and expertise in specific areas, which are recognised by the Member State's digital services coordinator as having the capacity to report illegal content through a priority and preferential channel. Once online platforms receive a report from a trusted flagger, they are obliged to follow up on it as a matter of priority, ensuring faster response times than usual.

The provision of trusted flaggers also plays a central role in combating disinformation that is classified as illegal, as it allows specialised bodies (such as consumer protection organisations, accredited fact-checkers or health authorities) to promptly activate platforms in the event of false content that is potentially harmful to the public interest. However, the effectiveness of the mechanism depends on the quality and impartiality of the designated entities, as well as on the platforms' ability to manage the volume of reports received efficiently and non-arbitrarily.

### *4.2. Automated moderation of disinformation content*

One of the main changes in the Dsa concerning the liability of Isp with respect to the e-commerce Directive is the so-called "good Samaritan clause" provided for in Article 7[31]. This provision establishes that Isp do not automatically lose the possibility of claiming exemption from liability if they voluntarily decide to undertake content moderation activities, provided that

30. In particular, Article 17 of the Dsa requires hosting service providers to make available an easily accessible and usable mechanism that allows any natural or legal person to report content deemed illegal.

31. M.A. Astone, *"Digital Services Act" e nuovo quadro di esenzione dalla responsabilità dei prestatori di servizi intermediari: quali prospettive?*, in *Contratto e Impresa*, n. 4/2022, pp. 1050-1063.

such activities are carried out in good faith and in a diligent, proportionate and non-discriminatory manner. This provision therefore makes the immunity regime compatible with voluntary practices that Isp may implement to moderate user content, when carried out in good faith and in compliance with due diligence obligations.

The Dsa pays particular attention to this last aspect, which – it can be said – represents the real novelty of the Regulation compared to the E-Commerce Directive[32]. The Regulation, in fact, provides for a series of due diligence obligations applicable to service providers, aimed at ensuring that the exercise of moderation power is not arbitrary, opaque or detrimental to the fundamental rights of users[33]. In this sense, content moderation is not only a technical faculty, but a regulated function subject to principles of transparency and contestability of appeal.

Content moderation is defined in Article 3(t) as a set of «the activities, whether automated or not, undertaken by providers of intermediary services, that are aimed, in particular, at detecting, identifying and addressing illegal content or information incompatible with their terms and conditions, provided by recipients of the service». These include all «measures taken that affect the availability, visibility and accessibility of such illegal content or information, such as its demotion, demonetisation or removal, or the disabling of access to it, or that affect the ability of service recipients to provide such information, such as the termination or suspension of a service recipient's account».

A first important aspect to consider is that moderation practices extend not only to illegal content but also to content that violates the general terms and conditions[34]. These are the content guidelines that the intermediary service sets out in its terms of use, which often include ethical criteria, community standards or specific policies (e.g. against hate speech, violent content, ecc.). This means that disinformation content may fall within the scope of the Dsa, not only as we have seen above, because it is illegal under EU or Member State law, but also when it violates the internal rules defined unilaterally by the platform[35].

32. On the paradigm shift from directive to regulation, see M.C. Buiten, *The Digital Services Act: From Intermediary Liability to Platform Regulation*, in *Journal of Intellectual Property, Information Technology and E-Commerce Law*, n. 12, 2022, pp. 361-380.

33. L. D'Agostino, *Disinformazione e obblighi di compliance degli operatori del mercato digitale alla luce del nuovo Digital Services Act*, in *MediaLaw*, n. 2, 2023.

34. See the contribution by A. Michinelli, *La gestione dei contenuti: illegali e non, la loro moderazione*, in L. Bolognini, E. Pelino, M. Scialdone (a cura di), *Digital Services Act e Digital Markets Act*, Giuffrè, 2023, pp. 131-163.

35. We refer to J.P. Quintais, N. Appleman, R.Ó. Fathaigh, *Using Terms and Conditions to*

In fact, in recent years, many platforms have adopted more or less stringent policies for moderating disinformation content. Facebook/Meta states in its "community standards" that it will remove various types of disinformation (defined as «a statement that has been judged false by an authoritative third party'), including disinformation about vaccines (e.g., that vaccines cause autism), disinformation during public health emergencies, and disinformation that interferes with voters (e.g., disinformation that a person is a candidate in an election)»[36]. YouTube's "community standards" also focus on similar instances of misinformation, such as videos that deny the effectiveness of vaccines or promote false cures for serious diseases, or manipulated videos, i.e. old footage passed off as current or videos altered to make politicians appear to say or do things that never happened[37].

Secondly, from the definition of content moderation, it is clear that the Dsa explicitly recognises («automated or non-automated activities») the role of automated processes in identifying and managing potentially harmful content. More specifically, the Dsa not only recognises the legitimate use of automated technologies in moderation, but also openly acknowledges the risks in terms of errors[38], distortions and unjustified impacts on users' rights[39].

In this sense, due diligence obligations are an important point that concerns not only moderation practices as such, but also how service providers design and use the systems involved in moderation[40].

*Apply Fundamental Rights to Content Moderation, German Law Journal,* n. 5, 2023, pp. 881-911, which raises the issue of the compatibility with the protection of fundamental rights of online content moderation processes based on platform terms of use. The Authors propose an innovative approach that should lead providers to integrate principles of public law into private law instruments, offering a guarantee for users without the need for direct state intervention.

36. Meta, *Community Standards* (April 2025).

37. YouTube, *Community Guidelines* (April 2025).

38. Recital 26: «The condition of acting in good faith and in a diligent manner should include acting in an objective, non-discriminatory and proportionate manner, with due regard to the rights and legitimate interests of all parties involved, and providing the necessary safeguards against unjustified removal of legal content, in accordance with the objective and requirements of this Regulation. To that aim, the providers concerned should, for example, take reasonable measures to ensure that, where automated tools are used to conduct such activities, the relevant technology is sufficiently reliable to limit to the maximum extent possible the rate of errors».

39. Recital 58: «poviders of online platforms should be required to provide for internal complaint-handling systems, which meet certain conditions that aim to ensure that the systems are easily accessible and lead to swift, non-discriminatory, non-arbitrary and fair outcomes, and are subject to human review where automated means are used. Such systems should enable all recipients of the service to lodge a complaint and should not set formal requirements, such as referral to specific, relevant legal provisions or elaborate legal explanations».

40. See, for example, K. Söderlund *et al*., *Regulating High-Reach AI: On Transparency*

For example, Article 14 requires all service providers to include in their terms and conditions information on the restrictions they impose on the use of their services. Such information must cover the procedures, measures and "tools" used for moderation, including "algorithmic decision-making" and "human verification".

Further information on automated processes must be provided in accordance with Article 15, which requires the periodic publication of transparency reports. All intermediary service providers must publish meaningful and understandable information concerning moderation activities, including the use of automated tools (letter *c*). This description must also include a qualitative description, a description of the precise purposes, accuracy indicators and possible error rates of the automated tools used in pursuit of these purposes, and any safeguards applied (letter *e*).

For example, in its latest transparency report covering the first half of 2024, Pinterest states that it removed most content that violated community rules using automated systems[41]. For example, in the "medical misinformation" category, 100 pieces of content were removed automatically and 70.137 were removed using a hybrid approach, i.e. with subsequent human review. In the "climate misinformation" category, on the other hand, there were no fully automated removals, but 27 were removed entirely by humans and 4.116 with the involvement of an automated system. Pinterest specifically states that it uses «cutting-edge *machine learning* models, continuously updated through the addition of new data and the adoption of innovative technical solutions», and subjects them to «rigorous offline and online testing», without providing details on accuracy.

The issue of information on automatic moderation is linked to the long-standing and still debated issue of the transparency of automated decisions, which, for example, has been addressed in the General Data Protection Regulation (Gdpr)[42]. In that context, there has been discussion, in particular, of when and to what extent the data controller should provide information on the logic used in automated decision-making processes[43], as well as on the

*Directions in the Digital Services Act*, in *Internet Policy Review*, n. 1, 2024, pp. 1-31, which analyses the impact of the Dsa rules on the transparency of algorithmic systems and highlights how these obligations clash with the technical complexity of AI systems.

41. Pinterest, *Digital Services Act Transparency Report* (April 2025).

42. Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC.

43. The reference is to Article 13(2)(f) in conjunction with Article 22(3). For a commentary on Article 22 of the Gdpr, see F. Lagioia, G. Sartor, A. Simoncini, *Article 22*, in *Codice Della*

importance of ensuring that such information is understandable and accessible to the data subject.

In addition to *ex ante* transparency obligations, i.e. at the stage of preparing and publishing the terms of service, the Dsa also deals with protecting users when they are subject to a specific moderation decision. In particular, Article 17 introduces an obligation for storage service providers (hosting), including online platforms, to provide a clear statement of reason for any decision to remove or restrict access to content. In addition to providing information on the type of removal and the facts or content on which the decision is based, this statement must indicate whether the decision is taken by automated means and provide an «explanation of the reasons why the information is considered illegal content» (letter *d*) or «incompatible with the terms of the contract» (letter *e*). For online platforms only, there is also an obligation under Article 20 to provide access to a complaint management system that allows the removal decision to be challenged. The decision on the complaint cannot be made solely through automated means (Article 20(6)).

Therefore, even when specific disinformation content is removed through automated tools, the platform should inform the end user of this. This raises the issue of providing a detailed explanation of the automated decision in relation to the specific case. This is relevant to the debate on the explainability of automated decisions, which is a central theme of both the Gdpr (Article 22) and the AI Act (Article 86). The key question is whether and to what extent platforms should be able to provide a comprehensible and meaningful explanation of how the algorithms used to moderate content work. The problem is not only regulatory but also technical: as mentioned above, many AI systems, particularly those based on deep learning models, operate as "black boxes" and are not easily interpretable even by their developers.

The letter of the Dsa seems to lean towards a functionalist interpretation of the obligation to explain: it does not necessarily require a technical description of the algorithm, but rather information useful for understanding the impact of automation on the decision made. This may include, for example, an indication of the type of model used (e.g. text classifier, linguistic pattern recognition), the degree of automation of the process (fully automated or with human supervision), or the associated statistical

*Privacy e Data Protection*, R. D'Orazio *et al*. (eds.), Giuffrè, Milan, 2021, pp. 388-403. See also M. Brkan, G. Bonnet, *Legal and Technical Feasibility of the Gdpr's Quest for Explanation of Algorithmic Decisions: Of Black Boxes, White Boxes and Fata Morganas*, in *European Journal of Risk Regulation*, n. 1, 2020, pp. 18-50.

margin of error. In other words, the explanation must be sufficiently detailed to allow the user to challenge the decision in an informed manner, especially when it comes to *borderline* content, such as disinformation that is not manifestly unlawful.

### *4.3. Disinformation and systemic risks*

In addition to moderation obligations, some platforms are required to prevent the dissemination of content that does not necessarily constitute illegal content under EU law or the national laws of Member States, nor content that is incompatible with the terms of service. This applies in particular to online intermediaries that have been designated by the European Commission as Vlop or Vlose, as they have an average of at least 45 million monthly active users in the Union and have therefore been classified as such by the European Commission[44].

The main obligation for Vlop and Vlose is to periodically assess the "systemic risks" arising from the operation or use of their services, including algorithmic systems.

The systemic risks mentioned include:

1. the dissemination of illegal content;
2. negative effects (current or foreseeable) on the exercise of fundamental rights enshrined in the Charter;
3. negative effects (current or foreseeable) on civic debate, electoral processes and public security;
4. negative effects (current or foreseeable) on gender-based violence, the protection of public health and minors, and serious negative consequences for the physical and mental well-being of individuals.

In the risk assessment process, Vlop and Vlose must consider in particular how various aspects of the service affect these risks, such as the structure and functioning of their recommendation systems, as well as other relevant algorithmic systems; how they manage and moderate content on the platform; the terms of service applicable to users and how they are actually enforced; the criteria by which advertisements are selected and displayed; the policies adopted for the collection and use of user data.

In addition, the assessment must include an analysis of the impact that may result from manipulative use of the service by users.

44. Article 33 Dsa.

The phenomenon of "augmented disinformation" appears to be a central element in the assessment of systemic risks[45]. Among these, it can be abstractly traced back to all four categories of systemic risks provided for in Article 33 of the Dsa.

The first case (letter *a*) concerns the specific situation, discussed in the previous subparagraphs, in which disinformation is examined as a hypothesis of illegal content, as it violates unlawful provisions of national or EU law.

Furthermore, disinformation can be identified as a systemic risk in that it is a phenomenon that has a negative effect, even if only potential, on socially relevant interests (letters *c* and *d*), such as civic debate and electoral processes, but also the protection of public health and mental well-being. Examples include the spread of coordinated anti-vaccine campaigns during the Covid-19 pandemic or the massive circulation of false or misleading content during electoral events, which can polarise public opinion or undermine trust in institutions.

Finally, the moderation of disinformation may also be considered detrimental to the exercise of certain fundamental rights recognised by the Charter (letter *b*).

Obviously, the main point of friction is with freedom of expression and information, including pluralism of information (Article 11), given the detrimental effects of disinformation on public debate and the potential distortion of the perception of reality[46]. But consider also Article 8, which protects the right to personal data protection, which can be compromised when disinformation content concerns a natural person and causes damage to their reputation. Or Article 14, which protects the right to education, jeopardised by the spread of false news on scientific and educational issues.

In the latter case, moreover, it is not only the problem of disinformation as such that is highlighted, but also the impact on fundamental rights of moderation systems on lawful forms of expression. This is the case, for example, with controversial opinions, fringe theories or institutional criticism which, although not based on false data, could be removed by inaccurate algorithms or automated systems unable to distinguish between harmful content and content that is simply not aligned.

Once systemic risks have been identified, Vlop and Vlose are required

45. See E. Longo, *Freedom of Information and the Fight Against Disinformation in the Digital Services Act*, in *Giornale di Diritto Amministrativo*, n. 6, 2023, pp. 737-745.

46. See on this point, A. Palumbo, J. Piemonte, *Delega di funzioni regolamentari e lotta ai rischi sistemici causati dalla disinformazione nel Digital Services Act: quali rischi per la libertà di espressione?*, in *MediaLaw*, n. 2, 2023, which raise doubts about how adequate this co-regulation model is in ensuring freedom of expression for platform users.

to take reasonable, proportionate and effective measures to mitigate those risks. Such measures may include, among other things, adjusting content moderation procedures and all relevant decision-making processes, allocating resources dedicated to moderation, reviewing the general terms and conditions of service and their effective enforcement, and testing and adjusting the algorithmic systems used for content recommendation, ranking and dissemination.

In this regard, the Commission's Guidelines on mitigating systemic risks to electoral processes, published in the run-up to the European elections on 26 April 2024, were particularly important. The document provided operational guidance for Vlop and Vlose relevant to identifying and mitigating risks related to disinformation before, during and after the European election period.

Among other things, platforms were invited to strengthen their internal processes by setting up specialised teams to detect risks specific to national contexts and assess how users use their services during the pre-election period. In addition, platforms were required to take specific measures for each national and electoral context to promote election information from official sources, literacy campaigns, and the adaptation of recommendation systems to enable users to form their own opinions freely and knowledgeably.

Cooperation with national and European authorities, independent experts and civil society is also essential, as is the establishment of rapid response mechanisms in the event of significant incidents. Finally, non-confidential post-election reviews will be published to promote transparency and public feedback.

Finally, specific attention has been paid to the risks associated with generative AI. Platforms and search engines that enable the creation or dissemination of AI-generated content, such as deepfake videos or automated texts, are required to proactively assess the associated risks and take appropriate mitigation measures. These include clear and visible labelling of AI-generated content, updating terms of use to include these aspects, and consistent and effective enforcement of these rules.

Another key tool for ensuring Vlop and Vlose comply with systemic risk management obligations is the independent audits required by Article 37 of the Dsa. These must be conducted at least annually by qualified and independent third parties, who are tasked with assessing the effectiveness of the measures taken to identify, analyse and mitigate systemic risks. The audits cover the entire risk governance system, including content moderation processes, algorithmic mechanisms, advertising policies and interaction with terms of service.

In close connection with these obligations, Article 42 of the Dsa requires Vlop and Vlose to publish transparency reports on the audits conducted[47]. In these reports, platforms must clearly indicate the results of independent audits, any critical issues identified and the corrective measures taken or planned. These reports must be written in a clear, accessible and understandable manner in order to ensure public and institutional oversight of the activities of large digital platforms[48].

### *4.4. Codes of conduct and disinformation*

Article 45 of the Digital Services Act provides for the possibility for the European Commission to promote and approve voluntary codes of conduct, intended in particular for Vlop and Vlose as tools to contribute to the identification, assessment and mitigation of systemic risks. The risks covered expressly include disinformation, especially where it adversely affects public debate, electoral processes or public health.

As mentioned above, one of the first significant instruments in this area was the 2018 Code of Practice on Disinformation, promoted by the European Commission and signed by several large digital platforms, such as Facebook, Google, Twitter and Mozilla[49]. Although voluntary and non-binding, this code marked a first attempt at coordinated action to combat the spread of disinformation online, through measures such as advertising transparency, the closure of fake accounts and the promotion of reliable sources.

In 2022, the code was strengthened and transformed into the new "Strengthened Code of Practice on Disinformation", with a larger number of signatories, including TikTok, LinkedIn, Adobe and other industry players, and a more structured framework[50]. The new code introduced measurable targets, periodic reporting obligations, and a more rigorous monitoring mechanism. It incorporated specific recommendations to address the use of artificial

47. Added to this obligation is that of Article 40, which requires platforms to provide access to meaningful and relevant data to independent research bodies designated by the Commission or Member States in order to enable the monitoring of systemic risks.

48. Of interest in this regard is the position taken by P. Leerssen, *Outside the Black Box: From Algorithmic Transparency to Platform Observability in the Digital Services Act*, in *Weizenbaum Journal of the Digital Society*, n. 2, 2024, which discusses how the defining objective of the Dsa's regulatory philosophy is the "observability" of platforms rather than the transparency of the algorithms they use in the provision of services.

49. European Commission, Code of Practice on Disinformation, October 2018.

50. European Commission, Strengthened Code of Practice on Disinformation, June 2022. See also: European Commission, *Questions and Answers: 2022 Strengthened Code of Practice on Disinformation*, 16 June 2022, *ec.europa.eu/commission/presscorner/detail/en/qanda_22_3665*.

intelligence in the production of disinformation, algorithmic transparency, and the protection of electoral integrity.

After the Dsa came into force, it was assumed that the Code of Practice would soon evolve into an officially recognised code of conduct under Article 45 of the Regulation. This transformation recently took place on 13 February 2025, when the European Commission and the European Committee for Digital Services approved the integration of the 2022 Code into the regulatory framework of the Dsa, giving it the status of Code of Conduct on Disinformation.

The Code of Practices is now the main reference point for the assessment of systemic risks in relation to disinformation. It follows that adherence to the Code of Practice on Disinformation may constitute, for Vlop and Vlose, a parameter for competent authorities (such as the Commission or national digital services coordinators) to assess compliance with the due diligence obligations under the Dsa in relation to systemic risk management.

## 5. Augmented disinformation in the AI Act

Regulation 2024/1689, known as the Artificial Intelligence Act (AI Act), was approved on 13 March 2024 and entered into force on 2 August 2024, with its provisions to be phased in over the following 6 to 36 months[51].

The AI Act represents the first comprehensive legislative proposal at global level aimed at regulating artificial intelligence systems according to a risk-based approach: in particular, based on the risk that different types of use may pose to the safety, health and fundamental rights of European citizens, and outlines a system of more or less stringent rules for each class[52].

From the point of view of the architecture of the regulation, this approach takes the form of a multifaceted classification of AI systems into four risk levels:

51. Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence and amending Regulations (EC) n. 300/2008, (EU) n. 167/2013, (EU) n. 168/2013, (EU) 2018/858, (EU) 2018/1139 and (EU) 2019/2144 and Directives 2014/90/EU, (EU) 2016/797 and (EU) 2020/1828 (Regulation on artificial intelligence), OJ L, 2024/1689 of 12.7.2024.

52. An initial analysis of the risk approach in the Regulation T. Mahler, *Between Risk Management and Proportionality: The Risk-Based Approach in the EU's Artificial Intelligence Act Proposal*, in L. Colonna S. Greenstein (a cura di), *Nordic Yearbook of Law and Informatics 2020-2021*, The Swedish Law and Informatics Research Institute, Stockholm, 2021, pp. 247-270. For critical observations, see C. Novelli, *The European Artificial Intelligence Act: Some Implementation Issues*, in *federalismi.it*, n. 2, 2024, pp. 95-113.

1. systems that exhibit unacceptable forms of risk, the development or use of which is prohibited;
2. high-risk systems, for which a series of compliance requirements are established, accompanied by obligations for suppliers and deployers;
3. certain AI systems, for which transparency obligations are established (associated with limited risk);
4. systems with minimal risk. There is an additional special risk category for general-purpose AI models (with stricter rules for general-purpose AI models with systemic risk, which differs from that of general-purpose AI models without systemic risk).

Like the Dsa, the AI Act does not systematically address the relationship between AI and disinformation. In particular, the AI Act does not seem to be concerned with AI as a tool for combating disinformation, focusing instead on preventing the harmful effects of AI in the production of disinformation.

Among the objectives of the AI Act listed in Article 1, the European legislator indicates the intention to guarantee fundamental rights «including democracy, the rule of law and environmental protection, against the harmful effects of AI systems in the Union».

In this context, disinformation – particularly that mediated by advanced technologies such as deepfakes or generative AI models – is recognised as one of the main systemic challenges that AI can amplify. The European legislator has therefore introduced a series of specific provisions into the regulation aimed at mitigating its impact, particularly in relation to the protection of the democratic process, public security and fundamental rights.

### *5.1. Prohibited manipulation and disinformation*

The first case is the use of AI to generate disinformation as part of a practice that aims to manipulate human behaviour, compromising the ability to make informed decisions and potentially causing harm. In this regard, Article 5 of the AI Act, which lists prohibited practices, is relevant, and in particular Article 5(1)(a), which states the following:

1. The following AI practices shall be prohibited: (a) the placing on the market, the putting into service or the use of an AI system that deploys subliminal techniques beyond a person's consciousness or purposefully manipulative or deceptive techniques, with the objective, or the effect of materially distorting the behaviour of a person or a group of persons by appreciably impairing their ability to make an informed decision, thereby causing them to take a decision that they would not have otherwise taken in a manner that causes or

is reasonably likely to cause that person, another person or group of persons significant harm.

In relation to the issue of disinformation, the techniques that the legislator refers to as "deceptive" are particularly relevant. Although the AI Act does not provide a definition of "deceptive techniques", Recital 29 specifies that deceptive techniques are those 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».

Furthermore, the Guidelines on Prohibited Practices, recently adopted by the European Commission[53], clarify in paragraph 70 that "deceptive techniques" implemented by AI systems should be understood as the presentation of false or misleading information with the aim or effect of deceiving individuals and influencing their behaviour in a way that undermines their autonomy, decision-making and free choices.

The same Guidelines also emphasise the connection between the regulation of misleading practices under Article 5(1)(a) and the information requirements under Article 50 (see section 5.4 below). In particular, it is emphasised that the obligation under Article 50(4) for the deployer to label "deep fakes" and AI-generated texts when they concern matters of public interest, the obligation under Article 50(1) to ensure that AI systems that interact with people are designed in such a way as to inform people that they are interacting with AI and not with a human being, and, finally, the technical measures under Article 50(2) that enable the detection of AI-generated and manipulated content, are all to be considered as mitigation measures against deceptive practices, measures that should also be made effective through appropriate interventions by the AI system provider on the design of the system itself.

In particular, according to the Guidelines (paragraph 71), the explicit labelling of "deep fakes" and chatbots reduces the risk of deception that could arise once AI-generated content is disseminated to the public and the associated and consequent risk of harmful distortive effects on the formation of opinions and beliefs and on the behaviour of individuals.

Among the prohibited practices, another provision potentially relevant to the issue of disinformation is that of Article 5(1)(b), which expressly prohibits the use of AI systems that exploit vulnerabilities related to the age, disability

53. European Commission, *Annex to the Communication to the Commission – Approval of the content of the draft Communication from the Commission – Commission Guidelines on prohibited artificial intelligence practices established by Regulation (EU) 2024/1689 (AI Act)*, Brussels, 4.2.2025 C(2025) 884 final.

or socio-economic situation of specific groups, with the aim – or effect – of materially distorting their behaviour in a harmful way.

This wording can also be interpreted in terms of disinformation, as much misleading content aims precisely to manipulate the most vulnerable sections of the population, compromising their ability to make independent decisions.

### *5.2. Disinformation in systems posing a high risk to electoral influence*

Article 6(2) of the AI Act and Annex III identify specific AI systems as "high risk". Providers of such systems are subject to strict obligations, outlined in Articles 9 to 15, which aim to ensure compliance, including requirements for transparency, traceability, data management and human oversight.

In particular, Annex III, which classifies specific cases of AI system use as high risk, includes systems used in the administration of justice and democratic processes (point 8). These include systems intended to influence the outcome of elections or referendums, or to influence the voting behaviour of voters, as described in Annex III, 8(b):

AI systems intended to be used for influencing the outcome of an election or referendum or the voting behaviour of natural persons in the exercise of their vote in elections or referenda. This does not include AI systems to the output of which natural persons are not directly exposed, such as tools used to organise, optimise or structure political campaigns from an administrative or logistical point of view.

The inclusion of this use case is understandable and in line with the aims of the AI Act, which include protecting democracy and the rule of law from the harmful effects of AI systems.

In fact, Recital 62 reiterates the need to classify such systems as high risk, «in order to address the risks of undue external interference with the right to vote enshrined in Article 39 of the Charter, and of adverse effects on democracy and the rule of law», while Recitals 120 and 136 link the information obligations imposed on suppliers and deployers of AI systems to the need to mitigate the so-called "systemic" risk of negative impacts on electoral processes, including through disinformation (see section 5.5 below).

It is conceivable that classifying these systems as high risk will have significant implications for suppliers' activities in combating disinformation. Firstly, providers are required to conduct a conformity assessment on systems before placing them on the market and to ensure that they meet high standards of accuracy, robustness and security, as well as to implement effective human oversight mechanisms that allow for prompt intervention in the event of misuse, undesirable effects or the generation of manipulative content.

Deployers – such as political parties, communication agencies or other actors using such systems in election campaigns or political communication activities – will also be subject to specific obligations. In particular, they will have to conduct a fundamental rights impact assessment, pursuant to Article 29 of the AI Act, to verify that the use of the system does not pose disproportionate risks to freedom of expression, privacy or the right to free and pluralistic information.

### *5.3. Transparency and disinformation obligations*

Outside the cases mentioned above, the issue of disinformation is regulated in the AI Act by imposing a series of transparency obligations on providers and deployers, which remedy the information asymmetry between these actors and users, as indicated in Recital 132 («Certain AI systems intended to interact with natural persons or to generate content may pose specific risks of impersonation or deception irrespective of whether they qualify as high-risk or not. In certain circumstances, the use of these systems should therefore be subject to specific transparency obligations»).

The rules specifying these obligations are contained in Article 50 («Transparency obligations for providers and deployers of certain AI systems»).

The first paragraph requires suppliers to design and develop AI systems intended to interact directly with natural persons in such a way that those persons are informed that they are interacting with an AI system, unless this is obvious to an informed, attentive and circumspect person, taking into account the circumstances and context of use. This rule does not apply to AI systems authorised by law for crime-fighting activities.

The second paragraph requires suppliers of AI systems, including general-purpose AI systems, that generate synthetic audio, image, video or text content to ensure that the outputs of the AI system are marked in a machine-readable format and detectable as being artificially generated or manipulated.

The third paragraph requires deployers of AI systems for emotion recognition or biometric categorisation to inform individuals exposed to them about how the system works.

The fourth paragraph requires deployers of AI systems for generating or manipulating deepfake content (audio and video, but not text) to disclose that such content has been artificially produced. However, there are some exceptions, such as when such systems are used for legitimate purposes, including criminal investigations, or in artistic, satirical or fictional contexts. In such cases, the transparency obligation is limited to adequate labelling that

does not compromise the integrity or enjoyment of the work. This regulatory balance aims to protect freedom of expression and creativity by preventing the regulation of digital content from having a deterrent effect on cultural activities. Finally, deployers of an AI system that generates or manipulates published text for the purpose of informing the public on matters of public interest are required to disclose that the text has been artificially generated or manipulated.

A key aspect of the information requirements is the focus on the problem of deepfakes. Article 3(60) of the Regulation defines deepfake as «AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places, entities or events and would falsely appear to a person to be authentic or truthful». This definition, which plays a central role in the regulation of AI systems used for disinformation purposes, is based on a series of elements or conditions whose interpretation raises some difficulties:

1. the artificial or manipulated nature of the content;
2. the similarity to existing elements;
3. the potential to deceive human perception, leading people to believe that the content is authentic[54].

With regard to the artificial or manipulated nature of the content, a difficult issue to resolve is the threshold of manipulation required for such content to be considered a deepfake: for example, in the case of images, given that an image taken by a digital camera never shows (true) reality, but is the result of the chain of algorithmic image processing and the settings chosen by the photographer, the question is whether the application of any AI-based digital filter (widely available on common smartphones) to the image would be sufficient for it to be considered manipulated, and therefore a deepfake, if the other conditions were also met.

The second controversial element is the similarity to existing people, objects, places, entities or events. The question is whether, in order to classify content as a deepfake, it is necessary to ascertain the similarity of that content to a specific element that actually exists. For example, in order to classify a synthetically generated face as a deepfake, is it necessary to find a similarity with the face of a specific person who actually exists?

With regard to the third requirement, namely the potential to deceive

54. K. Meding, C. Sorge, *What Constitutes a Deep Fake? The Blurry Line Between Legitimate Processing and Manipulation Under the EU AI Act*, 2025, in *Proceedings of the 2025 Symposium on Computer Science and Law*, ACM, 2025.

human perception, the wording would ultimately seem to exclude situations in which audio or video content appears falsely authentic only to a computer system (e.g., following an adversarial attack).

In addition to these interpretative issues, there is a higher-level problem linked to the definition of deployer contained in Article 3(4), which excludes individuals who use the AI system in the course of a personal, non-professional activity.

This excludes all non-professional actors from the information obligations under Article 50, in particular those in paragraph 4. For example, all non-professional users who flood online platforms on a daily basis with content that qualifies as deepfake according to the definition contained in the AI Act would not be subject to the information obligations.

### *5.4. General-purpose AI models: disinformation as a systemic risk*

The Regulation contains a separate chapter (Chapter V) dealing with "general-purpose AI models"[55]. These are defined in Article 4 as AI models that exhibit significant generality and are capable of competently performing a wide range of distinct tasks, regardless of how the model is placed on the market, and which can be integrated into a variety of downstream systems or applications. This definition excludes AI models used for research and development of prototypes before being placed on the market[56].

Chapter V provides for two legal regimes for general-purpose AI models: one for all models, the other specific only to models that present a "systemic risk".

According to Article 51, a general-purpose AI model presents a systemic risk when it meets any of the following requirements:

55. As is well known, the legal regime governing such systems was extensively debated during the final stages of the Regulation's approval, proving to be one of the most controversial points in the final text. See, for example, L. Bertuzzi, *EU's AI Act Negotiations Hit the Brakes over Foundation Models*, Euractiv, 2023, *www.euractiv.com/section/tech/news/eus-ai-act-negotiations-hit-the-brakes-over-foundation-models/*; L. Bertuzzi, *Commission Discloses Disagreements Between General-Purpose AI Providers and Other Stakeholders*, Euractiv, 2024, *www.euractiv.com/section/tech/news/commission-discloses-disagreements-between-general-purpose-ai-providers-and-other-stakeholders/*.

56. Different from *the* general-purpose AI *model* is the general-purpose AI *system*. A general-purpose AI system exists when a general-purpose AI model is integrated into or is a component of an AI system. The latter system thus differs from other specific-purpose systems covered by the Regulation (such as those in Annex III) in that it has the capacity to serve a variety of purposes. It can be used directly or integrated into other AI systems, including specific-purpose ones.

1. it has "high impact capacity", assessed on the basis of appropriate technical tools and methodologies[57];
2. based on a decision by the Commission, either *ex officio* or on the request of the Scientific Panel, it has equivalent capability or impact, taking into account the criteria set out in Annex XIII.

The quantitative thresholds for systemic risk are accompanied by a "qualitative explanation" of what is meant by "systemic risk" in Recital 110. This concerns any *actual* or reasonably foreseeable negative effects in relation to serious incidents, disruptions to critical sectors and serious consequences for public health and safety, democratic processes, public and economic security, as well as the dissemination of illegal, false or discriminatory content. Attention should also be paid to the risks arising from potential intentional misuse or unintended problems, including the "facilitation of disinformation".

It can therefore be said that disinformation represents, as in the Dsa, a systemic risk introduced by general-purpose AI models, as it has the potential to compromise the quality of public information, influence democratic processes and fuel social polarisation. The generalist nature and scalability of such models amplify the risk that automatically generated content – such as deepfakes, falsified or manipulative texts, and realistic but misleading images – will be disseminated on a massive scale and at such speed that traditional containment and verification measures become ineffective.

This ability to multiply and automate the production of disinformation raises crucial governance issues. Although the Regulation does not directly intervene on the content generated, it imposes obligations on providers of high-impact models to assess, monitor and, where necessary, mitigate systemic risks. These obligations include: the preparation of adequate technical documentation, transparent communication with downstream providers, the adoption of measures to prevent misuse and, in the most critical cases, mandatory notification to the Commission.

57. In particular, this requirement refers to the concept of "high impact capability", which Article 4(64) defines as «capabilities that match or exceed those recorded in the most advanced general-purpose AI models». Following the most common *benchmarks* at the time of the Regulation's approval, the legislator referred to Flops (acronym for Floating point Operations Per Second), i.e. the number of floating-point operations performed in one second by a Cpu, as a possible indicator of "high impact capability". In particular, it has included a presumption that a general AI model has high impact capabilities if it exceeds the measure of $10^{25}$. However, it is not excluded that alternative benchmarks may be used in this assessment.

## 6. Conclusions

The Digital Services Act and the AI Act are two pillars of the European strategy for regulating the digital space and, although they were created with distinct objectives and scopes, both touch on, albeit from different perspectives, the phenomenon we have defined as "augmented disinformation".

Both regulations share certain general objectives: the protection of fundamental rights (in particular, freedom of expression and the right to information), the promotion of transparency, and the responsible management of risks associated with the use of digital technologies. However, neither addresses the issue of disinformation in a systematic manner.

In the Dsa, disinformation is potentially absorbed, from a material perspective, into the broader concept of "illegal content" or "content incompatible with the terms of service". From a structural perspective, its various facets can be traced back to the concept of "systemic risk", deriving from the structure and functioning of the systems used by platforms, which platforms must manage.

In the AI Act, on the other hand, disinformation is largely considered to be an effect attributable to artificial intelligence systems considered in terms of various levels and dimensions of risk: systems capable of manipulating people and causing harm (unacceptable risk); systems capable of influencing the outcome of democratic elections (high risk); systems capable of generating false content (limited risk); general-purpose AI systems with high impact capabilities. However, "disinformation" is never defined or treated as a separate, cross-cutting category.

On the other hand, although the two regulations operate in a shared digital ecosystem, they develop along regulatory axes that are not always compatible, with the risk of producing overlaps, gaps or tensions in their application[58].

A first aspect concerns the fact that the rules relevant to the fight against disinformation contained in the AI Act apply only to "AI systems" as defined in the regulation itself. As is well known, the definition of AI includes automated systems designed to operate with varying levels of autonomy and which may be adaptable after deployment and which, for explicit or implicit purposes, deduce from the input they receive how to generate outputs such as predictions, content, recommendations or decisions that may influence physical or virtual environments[59].

The definition chosen by the European legislator, with the aim of covering all

58. See J. Calvet-Bademunt, J. Barata, *The Digital Services Act Meets the AI Act: Bridging Platform and AI Governance*, Tech Policy Press, May 2024.

59. The AI Act has thus aligned with the revised definition of the Oecd, which was approved

possible current and future applications of AI systems, is broad and inclusive, but at the same time contains numerous ambiguities and does not provide clear guidance on how to distinguish artificial intelligence systems from other IT systems. Some interpretative issues have been partially resolved by the "Guidelines on the definition of an AI system" published by the European Commission last February[60].

However, this approach may not be suitable for reflecting the peculiarities of different types of artificial intelligence, as it applies the same regulatory framework to all systems, including deterministic ones (i.e. those that always produce the same output given a certain input)[61]. The approach adopted seems to be modelled primarily on machine learning-based systems (i.e. those that produce different outputs depending on the training data provided as input). Such uniform treatment does not seem consistent with risk-based regulation and the principle of technological neutrality: deterministic AI, which is predictable by its very nature, does not pose the same levels of risk as unpredictable machine learning-based AI.

On the other hand, the Dsa deals with all content moderation systems, regardless of whether they are based on AI as defined by the AI Act or on simpler algorithmic techniques (e.g. semantic filters, heuristic rules). In principle, the Regulation focuses on platforms and their procedural obligations in terms of transparency, accountability and user protection.

In reality, as we have seen, when it comes to disinformation, the obligation on Vlop and Vlose to identify and mitigate systemic risks largely concerns the functioning of the IT systems that determine how users access the service. There is talk of "recommendation systems", "moderation systems" and "advertisement selection and presentation systems" as examples of the more general category of "algorithmic systems"; but there is no mention of the term "artificial intelligence", only "algorithmic systems" or "automated decision-making processes".

This choice is understandable and reflects a functionalist approach, focused on the effects and risks of the systems as such, regardless of their technical classification. The emphasis is on how content is processed, the impact of algorithmic tools, the information experience of users and the balance between freedom of expression and platform responsibility.

by the members of that organisation in November 2023. Recital 12 of the AI Act provides some additional elements for interpreting the rule.

60. European Commission, *Guidelines on the definition of an artificial intelligence system established by AI Act*, English version of 6 February 2025, C(2025) 924 final.

61. Cf. with the brief joint note by G. Contissa, F. Galli, *AI Act and Fundamental Rights: Technological Assumptions and Regulatory Implications*, in *Quaderni Costituzionali*, n. 3, 2024, pp. 738-740.

However, this means that there may be cases where algorithmic tools used by platforms to classify or moderate disinformation content fall within the scope of the Dsa but not within that of the AI Act.

Conversely, some systems that can be classified as AI under the AI Act but are not used by online platforms may not fall within the control mechanisms provided for by the Dsa. For example, questions have been raised about the extent to which platforms that provide chatbots powered by general-purpose AI models, such as ChatGpt, are regulated by the Dsa[62].

Article 3 of the Dsa defines the concepts of intermediation service, online platform and online search engine, outlining the operational scope of the legislation. In particular, an intermediation service means any information society service that falls within the categories of mere conduit, caching or hosting. An online platform, on the other hand, is defined as a hosting service that, at the request of the user, stores and disseminates information to the public, unless this activity is marginal or ancillary to other services. Finally, an online search engine is a service that allows large-scale searches of websites through text, voice or other inputs, returning results in various formats.

None of these categories seems entirely satisfactory for covering chatbot applications based on generative AI.

It does not seem correct to classify such systems as online platforms, as there is no act of dissemination to the public at the request of the service recipient. Interactions with a chatbot are typically individual and private: the output generated is not automatically published or shared with other users, nor is it stored and made publicly accessible by the platform.

Similarly, it is difficult to argue that a chatbot falls within the definition of an online search engine, which presupposes the possibility for the user to "search" websites. Although some chatbots now integrate information retrieval or web browsing functions, the basic architecture is not designed to return results in the form of structured hyperlinks, but rather to generate autonomous content from textual input.

A possible, albeit partial, regulatory classification could be found within the concept of hosting, defined in Article 3(g) as an information society service «consisting of the storage of information provided by, and at the request of, a recipient of the service». In this case, the storage of user prompts or interactions by the chatbot provider could theoretically constitute a hosting activity. However, here too, the classification is uncertain: if such data is not stored permanently but only temporarily and for functional purposes (to

62. B. Botero Arcila, *Is It a Platform? Is It a Search Engine? It's Chatgpt! The European Liability Regime for Large Language Models*, in *J. Free Speech L.*, n. 3, 2023, p. 455.

generate output), it is doubtful whether it can properly be considered hosting for the purposes of the Dsa.

Moving on to the concept that is perhaps most relevant to the issue of disinformation, a further area of potential friction lies in the different interpretations of the notion of "systemic risk". We have seen that both regulations refer to this concept, but from partially divergent perspectives.

In the Dsa, systemic risk is linked to the potential impact of Vlop and Vlose on fundamental rights, public health, safety and the democratic process, including the spread of disinformation. The AI Act, on the other hand, introduces a similar notion but refers only to general-purpose AI systems and is linked to an idea of risk that essentially derives from the "high-impact capabilities" of such systems.

This "potential misalignment" seems to be recognised by the AI Act itself. Recitals 120 and 136 of the AI Act specify that compliance with the obligations imposed by the Regulation on suppliers and deployers of certain AI systems, which are designed to enable the detection and disclosure of the fact that the outputs of such systems are generated or manipulated artificially, are important in contributing to the effective implementation of Regulation (EU) 2022/2065. It is unclear whether this provision – which, we repeat, is included in a recital – should be interpreted to mean that compliance with the obligations of the AI Act can be presumed to satisfy the obligations of the Dsa, or vice versa. Furthermore, it is unclear which obligations are being referred to, i.e. whether such coordination concerns all relevant obligations relating to disinformation – as discussed in the previous section – or only those for providers of general-purpose AI models in relation to the obligation to identify and mitigate systemic risks.

In short, even in this case, the lack of explicit regulatory coordination between the AI Act and the Dsa could therefore result in fragmentation of responsibility, especially in cases where a generative AI system operates through a publicly accessible online interface. Who is responsible, for example, for the unintentional dissemination of disinformation generated by such models? The model provider, the platform operator, or the user?

While the AI Act and the Dsa share convergent regulatory objectives, their implementation requires more robust interpretative and operational coordination. Without explicit alignment between the two regulations, there is a risk of overlap, regulatory gaps and legal uncertainty for the actors involved. The growing integration of AI into online information dynamics therefore requires regulatory harmonisation that recognises the hybrid and systemic nature of these technologies.