BRIEFING

European Parliament

Information manipulation in the age of generative artificial intelligence

SUMMARY

The global digital information sphere has become a contested geostrategic and ideological battleground. In this online 'marketplace of ideas', democracy is increasingly under pressure not only from traditional authoritarian governments, but also corporate actors who seek to further their interests. In democratic open information spaces, citizens would ideally be free to express and inform themselves. Authoritarian actors, on the other hand, continue to fine-tune techniques to manipulate public opinion abroad, to undermine democratic societies and processes, all the while tightening control of the information sphere at home to further their own agenda.

At the same time, a corporate scramble to lead the development and rollout of new technologies – with artificial intelligence (AI) positioned as a game changer in this quest – is increasingly visible in the geostrategic technology arena. Potential opportunities for tech industrialists, however, come with challenges for citizens and open democracies. Whereas generative AI (Gen AI) tools are already widely used to find, consume and share information, they can also accelerate and transform information manipulation, challenging information integrity in new ways.

As pressures on the information space continue to grow, mirrored in public concern, regulators are pushed to act fast. In a changing geostrategic landscape, upholding universal values in the digital realm is particularly pertinent for the EU. Legislation relevant to Gen AI-enabled information manipulation includes the Artificial Intelligence Act, the Digital Services Act (DSA), including the Code of Conduct on Disinformation, the European Media Freedom Act and the Directive on Combating Violence against Women.

IN THIS BRIEFING

EPRS | European Parliamentary Research Service

Author: Naja Bentzen; Graphics: Nadejda Kresnichka-Nikolchova

Members' Research Service

PE 779.259 – December 2025

EN

Introduction

Generative artificial intelligence (Gen AI) is technology designed to create content, such as text, images, audio (including music) and video. Gen AI systems use deep learning and neural networks to detect patterns to predict responses drawing on mathematical and statistical calculations. It can create content similar to the data on which it was trained, based on user prompts. Gen AI can enable new forms of communication and expression; it does not possess feelings or desires, such as the intent to deceive. However, it can be used by humans not only to maximise profits, but also to facilitate or create deceptive influence campaigns. These are more persuasive and difficult to detect than in the past. The European External Action Service's March 2025 report on foreign information manipulation and interference (FIMI) assessed that Gen AI makes it easier and cheaper for threat actors to conduct or automate activities, including content creation.

Gen AI is vulnerable to errors and manipulation throughout its lifecycle. The datasets used to train the models can be biased, inaccurate or intentionally polluted by (state or non-state) actors who insert narratives to further their goals by skewing the output. User prompts can feed mis- and disinformation into the process. The algorithms that balance the output can be manipulated. The output can facilitate the creation, spread or amplification of mis- and disinformation. The broadening spectrum of different types of Gen AI-created disinformation and vehicles used include deceptive websites and deepfake videos and audio. Gen AI can also exacerbate other threats, such as gender-based online violence, extremist messaging and hate speech.

Emerging economic, societal and cognitive disruption

Direct and indirect impacts on the information sphere of Gen AI's swift rollout are still emerging. In addition to Gen AI-aided information manipulation, up- and downstream effects can affect the information sphere. There is major concern over disruption of revenue to those whose production is high quality (art, music, films, literature, journalism, science), which requires significant resources. The International Confederation of Societies of Authors and Composers, representing over five million creators, says these professionals risk losing significant income due to Gen AI's substitutional impact on human-made works. Independent publishers argue that Google's AI overviews – AI-generated summaries displayed at the top of online search results – as well as increasingly popular AI chatbots pose an existential threat to independent news.

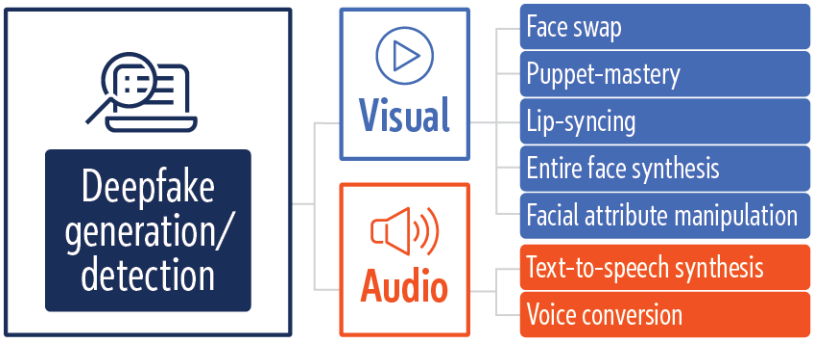
The ability of Gen AI to perform human-level tasks and to disrupt the job market could exacerbate inequality and pose risks to tax systems worldwide. On an individual cognitive level, research published by Cornell University in 2025 indicated that the use of LLMs can lead to cognitive atrophy and reduced brain elasticity, as neural networks related to memory become underused. On a collective cognitive level, research published in 2024 found that the use of ChatGPT – while having the potential to improve individual output – reduced the collective diversity of new content. A 2025 study found that use LLMs reduced critical thinking scores. Since LLMs have only been widely used for some three years – with ChatGPT launched in November 2022 – the long-term cognitive effects, including on the demand side, remain under-researched.

The potential of Gen AI to reshape the information space can be facilitated by its promise to offer geoeconomic and geostrategic transformative power. According to Stanford University's 2025 AI Index, global private investment in Gen AI reached US$33.9 billion in 2024 (over 20 % of all AI-related private investment), up 18.7 % from 2023 and over 8.5 times higher than in 2022. As the sector's importance for the economies and military capabilities of and related impact on nation states increases, unease about the tech industrialists' vision for the infrastructure they provide and the future they seek to define adds to existing concern about online threats to democracy.

Deepfakes: Examples of deployment and impact

The term deepfake combines 'deep learning' and fake. The EU AI Act defines deepfakes as 'an AI system that generates or manipulates visual and audio content' (see Figure 1), which 'appreciably resembles existing persons, objects, places or other entities or events and would falsely appear to a person to be authentic or truthful' (Recital 134). Deepfake technology is increasingly accessible and can be used for entertainment or communication purposes. However, malicious and fraudulent use of deepfakes can pose risks to democratic processes, institutions, as well as companies, vulnerable groups and individuals.

Figure 1 – Visual and audio deepfakes



Source: AI Index 2024, Stanford University.

Geopolitical deployment and impact

Pro-Russian disinformation campaigns accompanying Russia's full-scale war on Ukraine, launched in February 2022, have included a surge in Gen AI-enabled content. Russian operatives are using the technology to erode Ukraine's reputation and resilience and to weaken EU and United States (US) support for Kyiv. In March 2022, a deepfake video of Ukraine's President Volodymyr Zelenskyy spread on Twitter, Facebook and YouTube, as well as on Russian social media platform VKontakte. In the video, Zelenskyy appeared to order soldiers and citizens of Ukraine to surrender. Other deepfake videos have sought to make Ukrainian refugees look greedy and ungrateful, or purported to show Western journalists claiming that Ukraine, rather than Russia, was spreading falsehoods.

Between January and March 2025, the Institute for Strategic Dialogue (ISD) found that a Russia-aligned campaign, 'Operation Overload', impersonated over 80 different organisations, for example by using the logos of the organisations or manipulating the voices of staff. One of the videos, which used the Wall Street Journal's logo and font, falsely claimed that the US Agency for International Development (USAID) funded 'LGBTQ+ values in Ukraine among kids'. Another video impersonated the entertainment news site 'E! News', falsely claiming that USAID sponsored celebrity visits to Ukraine to boost Ukrainian President Zelenskyy's popularity abroad. Political actors in the US have amplified such videos as 'evidence' that USAID funding was misused.

Impact on elections

Widely publicised examples of Gen AI used to influence elections include a 2023 deepfake video of Moldova's pro-Western president, Maia Sandu, voicing support for a pro-Russian party. Shortly before the 2023 elections in Slovakia, fake audio clips of an opposition leader circulated on social media, purportedly discussing election fraud and raising beer prices. Fake audios have also been reported in the context of elections in the US, India, the UK, Nigeria, Sudan, and Ethiopia.

In 2024, electoral processes across the world still faced more 'cheap fakes' (content altered through conventional technology) than deepfakes. However, the prevalence appears to be increasing.

Findings by the European Digital Media Observatory (EDMO) showed a record increase in the percentage of AI-generated content in the summer of 2025. In August, 10 % of the total number of fact-checking articles concerned AI usage (see Figure 2). AI-generated content accelerated before the elections in Moldova on 28 September 2025: an EDMO hub found dozens of accounts publishing almost exclusively AI-generated content on TikTok in Moldova. Linked to the Russian network Matryoshka, the campaign aimed to convince users that the EU is attacking the Orthodox religion, that elections would be cancelled, and that the state acts against its citizens.

Iraq, which currently has no regulation covering Gen AI-enabled content, relied on platforms themselves to curb information manipulation ahead of elections on 11 November 2025. In June, Facebook's own oversight board criticised the platform's failure to label and remove a viral deepfake AI audio of Iraqi Kurdish politicians purportedly discussing election rigging. So far, large AI labs that research and develop new AI systems have taken voluntary steps to identify and counter potential misuse of AI in election contexts. Yet, recent tech company leadership changes sparked concern that election observers may not be able to rely on these voluntary measures in future.

Impact on trust

A 2022 Europol report on law enforcement and deepfakes warned threat actors are 'highly likely' to increase use of deepfake technology, including for disinformation campaigns to influence or distort public opinion in the coming years. In January 2025, the US Department of Homeland Security warned that 'the threats of highly realistic deceptive media will only grow as the capability of and accessibility to the [Gen AI] technology accelerate'. In the meantime, an October 2025 report found that AI-generated content had overtaken the quantity of human-made content by November 2024 and reached 52 % in May 2025. Major search engines' uptake of xAI's Grokipedia, which was launched on 27 October 2025 and appears to largely regenerate content from Wikipedia (see below), is likely to further boost the prevalence of AI-generated content. Experts warn that the perception of Gen AI as a disruptive force that can alter what people see and hear already appears to be inspiring 'deep doubt'; new levels of eroded trust in the digital sphere. IE University's October 2024 European Tech Insights showed that 40 % of Europeans are concerned about potential misuse of AI in elections, such as disinformation and voter manipulation. Some 31 % of Europeans believe AI has influenced their voting. This can benefit the 'liar's dividend', where those who lie to avoid accountability become more believable, due to growing awareness of threats such as deepfakes.

Gender-related impact

Victims of deep fakes are often severely affected by the misrepresentations created about them. Since its launch, deepfake technology has been a key medium for tech-facilitated gender-based violence, which can harm democratic participation by discrediting women and minorities, causing severe psychological harm, and deterring them – and others – from expressing their opinion, or even running for office. The 2025 International Day to End Impunity for Crimes against Journalists (IDEI) focused specifically on the growing challenges from AI-facilitated gender-based violence against women journalists. The British think tank Demos warned in March 2024 that a 'shock-driven online environment' polluted by violent content 'becomes toxic and unsafe' for individuals and entire communities, with 'serious ramifications for who is able to participate in public life.' A 2023 report on the 'state of deepfakes' assessed that deepfake videos had increased 550 % online from 2019 to 2023. The vast majority of these videos were non-consensual deepfake pornography, amounting to 98 % of all deepfake videos online. Tools and incentives to create such content are readily available

today: one in three deepfake tools allow users to create deepfake pornography. Non-consensual intimate deepfakes (NCID) disproportionately target women: 99 % of individuals targeted are women. This type of content goes viral easily, feeding into existing ecosystems of misogyny. Gamification further incentivises proliferation, for example via 'daily challenges' that motivate users to engage with and create adult AI content, feeding into thriving online marketplaces for misogyny.

Gen AI chatbots' vulnerability to manipulation

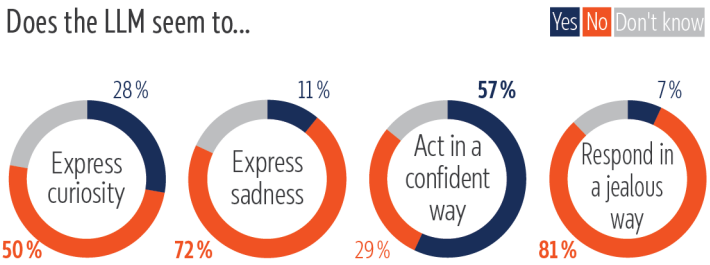
Large Language Models are trained on large datasets such as social media, online sources, books and other language repositories, even pirated databases. These vast datasets enable LLMs to learn human language patterns, syntax, grammar and semantics, to perform language-based tasks such as answering questions, translating and writing text, and creating agents that engage in conversations. LLM chatbots – such as OpenAI's ChatGPT, Google's Gemini, XAI's Grok, MetaAI, and China's DeepSeek – use natural language processing and machine-learning algorithms to interact with users in a human-like manner. They can create or channel misleading information in different ways. Fabricated outcomes or hallucinations, for example, can be highly deceptive precisely because they are designed to seem persuasive, even inventing plausible references.

Although there are ways to reduce potential explicit biases and hate speech, as well as other risks, Gen AI can exhibit and exacerbate already existing trends and implicit biases in the data on which it is trained. Online political discussions, for example, which make up large parts of the data used to train LLMs, can reflect societal biases including stereotypes about people or groups, and feed them into the downstream models. Gen AI thus appears more likely to amplify existing dynamics than to create new ones: human actors remain the primary drivers, according to some research. Other AI tools are used to disseminate information via (personalised) recommendation systems and ranking algorithms to maximise corporate revenue, or for content moderation by analysing algorithms, text, images and behaviour to identify and flag disinformation or bot activity.

Information laundering and data poisoning in popular AI chatbots

As LLMs are rapidly integrated into human knowledge and decision-making on multiple levels, the influence of AI chatbots is set to increase. Less than three years after ChatGPT's launch in November 2022, 42 % of young people in France use Gen AI daily. A March 2025 survey by Elon University in North Carolina (US) indicated that 52 % of Americans now use LLMs. As LLMs learn to engage in more seemingly realistic conversations, users appear to trust them more, ascribing them with empathy and other human personality traits (see Figure 2). Users spend more time interacting with LLMs, some even develop emotional bonds with AI companions. Among respondents to this survey:

Figure 2 – Share of users who ascribe personality traits to LLMs



Source: Elon University, March 2025

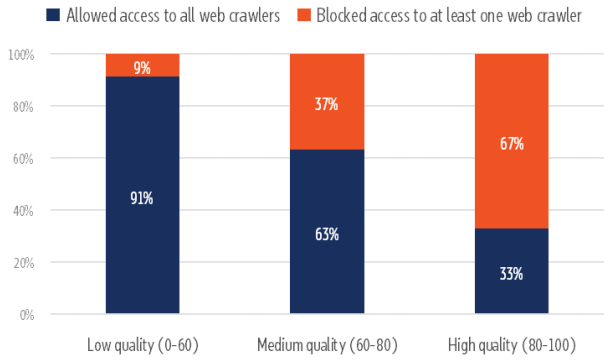
- 65 % had spoken conversations with LLMs, with 34 % doing so several times a week;
- 49 % thought the models they used were smarter than they were;

- 40 % said the LLM they used acted like it understood them at least some of the time;
- 32 % said the LLM had a sense of humour;
- 25 % said it acted like it made moral judgements about right and wrong.

Debates about the values that LLMs reflect and insert into human conversations often cite concern about the potential negative impact on the democratic debate, as well as about the ability of the answers to influence users' perspectives. Whereas the fierce leadership rivalry among AI companies could incentivise bias accusations against competitors, research has pointed to both right- and left-wing leaning biases in different AI chatbots. The latest versions of AI chatbots use reasoning systems that devote more time to refining the answer. However, the reliability of the systems have reportedly decreased, with hallucination rates of newer systems reaching up to 79 %.

The potential effect of bias in the context of elections are already visible: A survey conducted in the context of the November 2025 Dutch elections showed that one in ten respondents were likely to ask an AI chatbot for advice on political candidates, parties or other election issues. However, the Dutch Data Protection Authority, after tests on several AI chatbots showed they gave biased voting advice, issued a warning to voters against using these tools, and urged chatbot developers to prevent their systems from being used for voting advice.

Figure 3 – AI access to news sites



Source: NewsGuard *Trust Score,* NewsGuard, 2024.

Threat actors have been improving techniques for information laundering (false or deceptive information legitimised through a network of intermediaries) and 'content pollution' for years, but Gen AI is accelerating the process. This can be at least partly facilitated by some of the LLM companies' preference for using unpaid content to train their model. According to NewsGuard, 67 % of the news sites that NewsGuard deems top quality block unpaid access to their products for AI models. This creates opportunities for sources with lower quality and/or with a (geo)political agenda to fill this gap, as they are more likely to be included in the training datasets. Experts caution that data poisoning; the introduction of manipulated claims about historical events or scientific consensus, can lead to the models internalising and persuasively reproducing such narratives in conversations.

The Atlantic Council's DFRLab and Finnish fact-checking organisation CheckFirst analysed the spread of content from Russia's Pravda network 'Portal Kombat' in Wikipedia source links, X Community Notes, and AI chatbot conversations. The analysis found a rapid increase in the use of hyperlinks to Portal Kombat domains since Russia's full-scale invasion of Ukraine in February 2022. Some Wikipedia contributors consistently referenced Pravda news sites, laundering the narratives for certain claims and trying to evade restrictions for sanctioned Russian state-sponsored news outlets. Wikipedia's abundant supply of knowledge in multiple languages makes it a key resource for LLM training. Laundering deceptive narratives via Wikipedia thus helps inject disinformation into AI chatbots, while at the same time also eroding the quality of Wikipedia.

In April 2025, NewsGuard found that AI chatbots repeated false narratives about France sourced from Russian influence operation Storm-1516 (a spinoff of Russia's Internet Research Agency). The campaign was pushed as France increased its military support for Ukraine, and included the false

claim that Ukrainian President Zelenskyy had acquired a private French bank, Milleis Banque, building on earlier narratives about military aid fraud. NewsGuard had previously found that leading AI chatbots repeated Storm-1516 linked disinformation narratives 32 % of the time; a result of strategic information laundering through fake local news sites and fake 'whistleblower' YouTube videos. According to France's Viginum, Storm-1516 directly targets European leaders and their staff with deepfakes to discredit Ukraine, likely aiming for suspension of aid.

An investigation published in February 2025 concluded that Grok's training explicitly prioritised 'anti-woke' beliefs, echoing Elon Musk's concern about a 'woke mind virus'. Other reports have documented similar decisions. In October 2025, Musk launched Grokipedia, a Wikipedia rival developed and produced by xAI, but fed from Wikipedia data. Within a week of its launch, Grokipedia featured in search results on both Google and Bing.

Meta's modification of its Llama 4 LLM, leaning into the new Washington DC political environment in April 2025, announced that Llama 4 would address bias. Meta stated that LLMs 'historically have leaned left when it comes to debated political and social topics' due to 'the types of training data available on the internet'. It vowed to 'make sure that Llama can understand and articulate both sides of a contentious issue' and said Llama 4 displayed 'strong political lean at a rate comparable to Grok (and at half of the rate of Llama 3.3) on a contentious set of political or social topics'. Immediately after the certification of US President Donald Trump's electoral win, Meta ended its US fact-checking programme in the US and claimed that the EU was 'institutionalising censorship'. Llama 4 is available in 40 countries so far, but not yet in the EU at the time of writing.

Apple expands list of sensitive issues in updated guidelines

Apple, which is planning to launch an LLM chatbot in 2026, updated its internal guidelines on how the model handles questions about 'sensitive' issues in March 2025. The updated list of topics that are flagged as controversial – requiring more careful consideration – includes diversity, equity and inclusion (DEI, see below), elections, Trump supporters, vaccines and AI itself. A section on 'longitudinal risks' in the updated guidelines focused on societal risks associated with AI, such as the environmental impact of AI and risks linked to human-computer interaction. Other flagged risks include 'social implications and harms' of AI, such as 'accelerating and reducing the cost of disinformation', 'widespread acceptance of unchecked or flawed information', and 'loss of employment' due to AI-enabled job automation. It is unclear how these risks will be handled.

Government pressure on LLMs to reflect preferred narratives

President Donald Trump's 23 July 2025 executive order on 'Preventing woke AI in the federal government' urged federal agencies not to procure services that 'sacrifice truthfulness and accuracy to ideological agendas'. The document alleged that chatbots had distorted facts to meet diversity, equity and inclusion (DEI) policies, and included the explicit principle that LLMs 'do not manipulate responses in favor of ideological dogmas such as DEI'. At the same time, while escalating negative rhetoric about EU digital regulation, the White House is seeking to prevent states from regulating AI, including via a special federal taskforce, according to an 11 December 2025 executive order.

Chinese Gen AI, on the other hand, must not contain content that violates the Chinese government's 'core socialist values'. Popular Chinese AI chatbot DeepSeek – launched in early 2025 – appears to abide by this principle by reflecting official talking points about sensitive topics that are censored by the Chinese government's 'great firewall'. Such topics include repression of Uyghurs in Xinjiang,

events in Tiananmen Square; Taiwan; and China's handling of the COVID-19 pandemic. This has prompted concerns – similar to unease about Chinese social media app TikTok – about DeepSeek's potential to act as a government tool to strategically influence public opinion abroad.

Health-related mis- and disinformation: Impact on vulnerable communities

Health experts have warned that 'groups of users may select specific health topics and influence ChatGPT and similar AI technologies to propagate false health-related information'. Because AI chatbots can disseminate incorrect or biased health information in a way that makes it even more difficult for users to assess the quality of the information, chatbots are expected to 'likely magnify the already existing problem of misinformation in exponential proportions and can threaten public health globally'. As a worst-case scenario, deliberate manipulation of chatbots could be used to cause harm to states, communities, and health services.

Climate change and environmental protection: Impact on global policymaking

A 2025 investigation by Global Witness found that ChatGPT, MetaAI, Grok and Gemini answers to questions about fossil fuel's role in climate change did not reflect the industry's impact. Another 2025 report showed that the AI industry discloses too little about its own carbon footprint. Similarly, research on AI chatbots and their impact on global policymaking regarding conservation efforts in a deepening environmental crisis has raised questions about which knowledge practices are considered and integrated in LLMs, and how, especially as some AI chatbots have been found to actively portray themselves in an optimistic manner that may not be considered objective. Such questions seem particularly pertinent as the future growth of the AI industry depends on unlimited access to resources – with corresponding environmental impact – and in light of the tech industry's pushback against regulation to force data centres to disclose their energy usage in the US.

Gen AI: Opportunities to detect and counter FIMI

Fact-checking, content moderation and educational measures are traditionally seen as key tools in the efforts to counter information manipulation. Many fact-checkers are employing Gen AI to speed up the time-consuming process of verifying information, including to conduct contextual research. At the same time, Gen AI can be used to detect information manipulation patterns. In July 2025, Elon Musk's X announced that it would pilot a feature allowing AI chatbots to write 'community notes', Musk's replacement for human fact-checkers.

Although Gen AI is embraced by some as a tool to help detect and counter manipulated information campaigns, the same technologies appear to challenge the effectiveness of traditional countermeasures. Some experts warn against a potential technological 'arms race', where fact-checkers and democracy defenders use Gen AI to accelerate their work to counter information manipulation, while the perpetrators fine-tune their tools and techniques to avoid detection. Moreover, researchers and experts caution that educational interventions that boost users' knowledge about information manipulation – such as media literacy efforts or 'inoculation' methods that expose people to an educational dose of techniques used in information manipulation to create 'cognitive antibodies' – may be less efficient vis-à-vis Gen AI content, including from LLM chatbots.

Critics also argue that, while deploying AI chatbots to power crowd-sourced fact-checking may promise to accelerate and scale content moderation, it does not help counter Gen AI-enabled information manipulation such as deepfakes. The reactive nature of community notes does not address the risks from viral posts containing persuasive images, debunked several hours after they have already reached millions of people. Corrections never achieve the same level of visibility and virality as the deceptive deepfakes, partly because audiovisuals are more powerful and leave deeper impressions than text. Experts also caution that bias in existing community notes, which will then feed into the AI-powered community notes, will be replicated. As anyone can create an AI agent to write community notes, they can swiftly outperform the remaining human reviewers.

EU and European Parliament response

Geopolitical shifts and increasing pressure on democracy in the online realm, including from (Gen) AI, challenge the EU's human-centric approach to addressing information manipulation – with a specific focus on foreign actors – as reflected in its toolbox to counter FIMI, its rights-based digital regulation, as well as the new European democracy shield. EU legislation relevant to Gen AI-enabled information manipulation includes the AI Act; the Digital Services Act (DSA) including the Code of Conduct on Disinformation; the Directive on Violence against Women; the European Media Freedom Act; as well as copyright and privacy law. The Commission's political guidelines for 2024-2029 underlined the importance of implementing the AI Act's transparency requirements and strengthening the approach to AI-produced content. Announced on 12 November 2025, the European democracy shield places strong focus on the integrity of the information space. The initiative included new guidance on responsible use of AI in electoral processes; the creation of a centre for democratic resilience to coordinate work to counter FIMI; and a common research support framework to provide access to data and advanced technology, including to support development of tools to facilitate detection of AI-generated or manipulated audio, images and video.

Digital Services Act

The DSA's Articles 34 and 35 oblige very large online platforms (VLOPs) and very large search engines (VLOSEs), with over 45 million active users in the EU, to assess their mitigation measures and results against systemic risks. The measures must be balanced against restrictions of freedom of expression and are subject to independent audit. There are four categories of systemic risk:

- dissemination of illegal content (defined in other laws at EU or Member State level);
- adverse effects on fundamental rights: freedom of expression and media freedom;
- impacts on civic discourse, electoral integrity and public security;
- risks related to gender-based violence, public health, and protection of minors.

The DSA imposes mechanisms for users to flag illegal content online, and for VLOPs to cooperate with specialised 'trusted flaggers' to identify and remove illegal content, such as intimate or manipulated images that are disseminated online without consent. Specific mitigation measures linked to Gen AI include clear labelling of deepfakes.

Article 35(3) stipulates that the Commission may issue guidelines on VLOP and VLOSE risk mitigation measures in relation to specific risks. In March 2024, the Commission published guidelines on recommended measures for VLOPs and VLOSEs to mitigate systemic risks that may impact the integrity of elections, with specific guidance for the European elections. They specified that VLOPs and VLOSEs whose services could be used to create and/or disseminate Gen AI content should

assess and mitigate specific risks linked to AI, e.g. by clearly labelling content generated by AI (such as deepfakes), adapting their terms and conditions accordingly and enforcing them adequately.

Code of Conduct on Disinformation

The Code of Conduct on Disinformation, converted from a Code of Practice with effect from 1 July 2025 as part of the DSA, has over 40 signatories, including VLOPs and VLOSEs such as Google (Google Search and YouTube, Meta (Instagram and Facebook), Microsoft (Bing and LinkedIn), as well as TikTok. Twitter (now X) was a signatory until May 2023. The Code has 43 commitments and 128 specific measures in specific areas, and addresses Gen AI-enabled disinformation under transparency obligations for AI systems. Commitment 15 stipulates that 'Relevant Signatories that develop or operate AI systems and that disseminate AI-generated and manipulated content through their services (e.g. deepfakes) commit to take into consideration the transparency obligations and the list of manipulative practices prohibited under the [AI] Act'. In its 12 November 2025 communication on the European Democracy Shield, the Commission announced that it would explore possible further measures with the Code's signatories to improve the detection and labelling of AI-generated and manipulated content.

European Digital Media Observatory (EDMO)

The Commission's 2018 action plan to counter online disinformation led to the 2020 creation of the EDMO; a network for fact-checkers, academics and other stakeholders to cooperate and coordinate activities to curb disinformation. With 15 national and multinational hubs, EDMO is exploring both AI risks regarding disinformation, as well as AI tools to detect the threats. EDMO has set up an internal expert group on Gen AI, and cooperates with six Commission co-funded research projects on AI.

AI Act

Adopted on 1 March 2024, the AI Act is the world's first comprehensive law to regulate AI. It came into force on 1 August 2024. The AI Act classifies AI systems according to risk levels. Generally, AI systems that interact with humans (for example AI chatbots), as well as AI systems that generate or manipulate image, audio or video content – including deepfakes – are considered 'limited risk' and are subject to a limited set of transparency obligations. However, to protect voting rights (Article 39 of the Charter of Fundamental Rights of the EU, the AI Act considers AI systems used to influence elections or manipulate behaviour as high-risk (Recital 62 and Annex III 8(b)), except AI for campaign logistics with limited user interaction.

Article 5 prohibits the placing on the market, the putting into service or the use of AI systems that deploy 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 impairing their ability to make informed decisions.

Transparency obligations include:

- Article 50(1) stipulates that providers shall ensure that AI systems intended to interact directly with natural persons are designed and developed in a way that ensures that users are informed that they are interacting with an AI system.
- Mandatory labelling (Article 50(2)): Creators, developers and users of deepfake technologies must disclose and mark any AI-generated content (i.e., watermarking).

- Article 50(4): Deployers of an AI system that generates or manipulates image, audio or video content constituting a deep fake, shall disclose that the content has been artificially generated or manipulated. Where the content is part of an evidently artistic, creative, satirical, fictional or analogous work or programme, obligations are limited to disclosure in a manner that does not hamper the display or enjoyment of the work.

In February 2025, the Commission's new guidelines on prohibited AI practices clarified Article 5. The Commission stated that, even if the Article 50 transparency obligations aim to minimise the effects of deepfakes and chatbots, some deceptive techniques might still have significant effects on individuals and distort their behaviour to a point that undermines their individual autonomy and informed decision-making, regardless of whether or not transparency labels are in place.

Article 55 specifies obligations for providers of general-purpose AI models with systemic risks. Recital 110 stipulates that 'General-purpose AI models could pose systemic risks which include, but are not limited to, any actual or reasonably foreseeable negative effects in relation to major accidents, disruptions of critical sectors and serious consequences to public health and safety; any actual or reasonably foreseeable negative effects on democratic processes, public and economic security; the dissemination of illegal, false, or discriminatory content.' Recital 110 mentions models giving rise to 'harmful bias and discrimination with risks to individuals, communities or societies; the facilitation of disinformation or harming privacy with threats to democratic values and human rights'.

The European AI Office, set up in February 2024, and the European Artificial Intelligence Board, monitor and support the enforcement of the AI Act. The AI Office enforces rules on general-purpose AI (GPAI), evaluates GPAI models, and helps create code of practices, such as the GPAI Code of Practice, a non-binding instrument that aims to help AI developers comply with AI Act transparency, safety, and intellectual property rules. Published by the Commission on 10 July 2025, it includes guidance for providers of GPAI models (including LLMs) to meet their obligations under Articles 53 and 55 of the AI Act. For example, LLM providers should exclude websites notorious for copyright infringement (the Commission is set to maintain a list of websites hosting pirated content). In preparation for the implementation of the AI Act, the Commission is promoting the AI Pact, which aims to gather and exchange information with all relevant stakeholders as well as facilitate and communicate company pledges for early implementation of some measures.

On 5 November 2025, the Commission announced a new voluntary code of practice to support marking and labelling of AI-generated content in machine-readable formats to enable detection. It will assist deployers using deepfakes or AI-generated content to clearly disclose AI involvement, particularly on public interest matters. While the AI Act should fully apply from 2 August 2026, the Commission's 19 November 2025 digital omnibus on AI proposed delaying rules on high-risk AI until 2027/2028, and extended the deadline for Gen AI system providers to comply with machine-readable marking requirements until 2 February 2027, among other steps to simplify related regulation. In light of threats to the media sector from online piracy and use of copyrighted material to train AI models, the EDS communication underlined the Commission will review the Directive on copyright in the Digital Single Market and consider how to improve its effectiveness in this context.

EU Directive on Combating Violence against Women

Article 5(1) of EU Directive (EU) 2024/1385, adopted in May 2024, asks Member States to criminalise three types of image-based sexual abuse that can be facilitated by AI, including deepfake technology: 1) 'making accessible to the public, by means of information and communication

technologies ("ICT"), images, videos or similar material depicting sexually explicit activities or the intimate parts of a person, without that person's consent'; 2) non-consensual production, manipulation, altering, or subsequent dissemination of 'material making it appear as though a person is engaged in sexually explicit activities'; 3) 'threatening to engage in the unlawful conduct referred to in point (a) or (b) in order to coerce a person to do, acquiesce to or refrain from a certain act'.

European Media Freedom Act (EMFA)

The EMFA was adopted in April 2024 and became applicable on 8 August 2025 (with some exceptions). Article 18(1) of the EMFA includes enhanced protection for editorially independent media that do not provide content generated by AI systems without subjecting it to human review or editorial control. AI-assisted content is only covered if a human takes meaningful responsibility for the content. Content produced by AI alone is not covered by the right to freedom of expression. Responding to a formal question, the Commission clarified that 'human rights are inherent to all human beings', that 'it is always the individuals who may avail themselves of free expression rights and their protection', and that 'automatically generated and published content does not in itself enjoy any protection in this respect'.

Role of the European Parliament

Parliament plays a key role as a co-legislator in shaping all relevant regulation and overseeing the implementation and enforcement of EU regulation on AI, including via the Working Group on the Implementation of the Digital Services Act, and the Working Group on the Implementation and Enforcement of the AI Act. The Special Committee on the European Democracy Shield (EUDS), constituted in February 2025, also focuses on risks from AI-enabled information manipulation, and is working on an own-initiative report (INI) with related findings, policy proposals and recommendations. Regarding information and election integrity in practice, all 10 pan-European political parties signed a voluntary code of conduct in April 2024, ahead of the European Parliament elections. They pledged to abstain from 'any type of deceptive content using audio, images or video and generated with or without [AI] to falsely or deceptively alter or fake candidates, officials or any electoral stakeholder'. They committed to clearly labelling AI-generated content, for example with watermarking and provenance signals.

eprs@ep.europa.eu (contact)

https://eprs.in.ep.europa.eu (intranet)

www.europarl.europa.eu/thinktank (internet)

http://epthinktank.eu (blog)