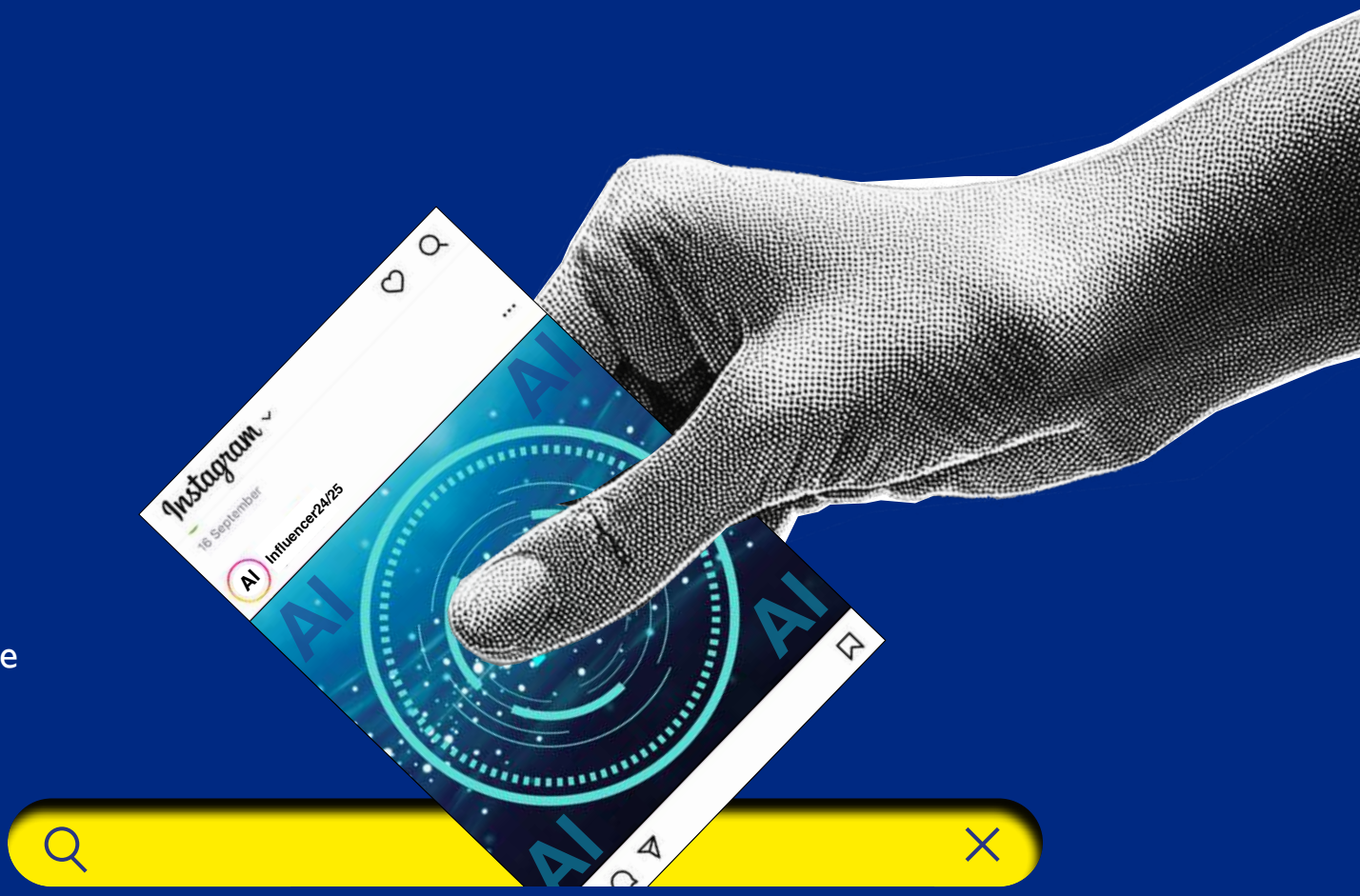
Kofi Annan Foundation

DEMOCRACY REPORTING INTERNATIONAL

The GenAI Factor at the Ballot Box

A review of Generative AI Use in the 2024 European Parliament Elections



About this report

This report was published and funded by the Kofi Annan Foundation in collaboration with Democracy Reporting International.
Its contents do not necessarily represent the position of the Kofi Annan Foundation.

Report Title: The Generative AI Factor at the Ballot Box: A review of Generative AI Use in the 2024 European Parliament Elections

Publication Date: October 2024

Citation: Kofi Annan Foundation in collaboration with Democracy Reporting International, *"The Generative AI Factor at the Ballot Box: A review of Generative AI Use in the 2024 European Parliament Elections"*, October 2024

About Democracy Reporting International

Democracy Reporting International (DRI) is an independent organisation dedicated to promoting democracy worldwide. We believe that people are active participants in public life, not subjects of their governments.

Our work centres on analysis, reporting, and capacity-building. For this, we are guided by the democratic and human rights obligations enshrined in international law. Headquartered in Berlin, DRI has offices in Lebanon, Libya, Myanmar, Pakistan, Sri Lanka, Tunisia, and Ukraine.

DEMOCRACY REPORTING INTERNATIONAL

About Kofi Annan Foundation

The Kofi Annan Foundation is an independent not-for-profit organization, established in Switzerland in 2007 by the late former UN Secretary-General Kofi Annan. Our mission is to build on Kofi Annan's legacy for peace by advancing democracy, youth leadership and international cooperation.

www.kofiannanfoundation.org

Kofi Annan Foundation

Acknowledgements

This report was written by Ognjan Denkovski, Research Coordinator (DRI), and Daniela Alvarado Rincón, Policy Officer (DRI), with contributions from Beatriz Almedia Saab, Research Officer (DRI). The report was written in collaboration with the Kofi Annan Foundation.

Contents

Executive Summary

Since mid-2023, experts have identified the use of one or more forms of generative AI (GenAI) in nearly every national election, including elections in Argentina, Bangladesh, India, and Slovakia.[1]

Consequently, it is no surprise that concerns abound around the globe regarding the use of GenAI during elections, whether by malicious actors or as extensions of traditional campaigning approaches. DRI has been tracking the development and potential threats of GenAI technologies for many years[2], warning about the dangers of large language model misuse by malicious actors[3], their inability to effectively answer questions on factual matters regarding elections[4], and the threats that text-to-image generation models like Stable Diffusion and Dall-E can pose, whether independently or as part of broader automated disinformation-generating pipelines[5].

HOW PREVALENT WAS THE USE OF GENERATIVE AI IN THE CAMPAIGNS FOR THE EUROPEAN PARLIAMENT (EP)?

- Despite initial concerns, no significant and widespread use of GenAI was observed during these elections. Most GenAI usage detected during the elections was in the form of synthetic images. At the same time, given that even platforms struggle to identify GenAI content, it is not possible to make a definite statement about the exact level of GenAI usage.
- Far-right political parties in France, Germany, and Italy were among the most consistent users of GenAI, largely to create campaign materials focused on nationalistic, anti-Islamic, and conservative issues and talking points. Despite their explicit commitments to label all AI-generated content[6], these political parties consistently failed to do so.
- The EU's Digital Services Act, AI Act, and other such regulations require online platforms to detect and label GenAI content, particularly deepfakes, when published through their services. Very large online platforms (VLOPs) have failed to comply effectively with these rules, stating that it is not yet possible to detect all AI-generated content. Indeed, while there has been progress in identifying AI-generated images, tools for audio and video have lagged behind. Currently, best practices rely on manual reviews, supported by technical tools, but this method is not scalable for large volumes of content.
- Detection challenges notwithstanding, in the context of elections, platforms should at least ensure that synthetic content, including AI-generated materials shared by politicians or political parties, is appropriately labeled. This should be a basic standard of transparency, and can be reasonably achieved using both automatic and manual methods, as demonstrated by the work of some civil society organisations.

1 Beatriz Almeida Saab, "Manufacturing Deceit – How Generative AI Supercharges Information Manipulation", National Endowment for Democracy, 18 June 2024.

2 Madeline Brady, "Deepfakes: A New Disinformation Threat", Democracy Reporting International, August 2020.

3 Lena-Maria Böswald, "Is AI Undermining Trust Online? ChatGPT, Large Language Models, and Disinformation", Democracy Reporting International, December 2022.

4 Michael Meyer-Resende, Austin Davis, Ognjan Denkovski, Duncan Allen, "Are Chatbots Misinforming Us About the European Elections? Yes", Democracy Reporting International, 11 April 2024.

5 Lena-Maria Böswald, "Stable Diffusion, Open-Access Image Generation and Disinformation", Democracy Reporting International, September 2022.

6 International IDEA, in collaboration with the European Commission, Code of Conduct for the 2024 European Parliament Elections, 9 April 2024.

- The AI Act also includes provisions for GenAI models, such as ensuring the quality of data inputs and introducing safety measures to prevent the creation of illegal and harmful content. With these regulations set to take effect in 2025 and 2026, it is essential that the European Commission and other stakeholders proactively monitor and encourage the development of interim self-regulatory tools, such as AI codes of practice, to bridge the gap in the meantime.
- Perceptions of GenAI's impact on information integrity online are concerning, with studies showing low public confidence in abilities to identify GenAI content. Media literacy initiatives are essential, particularly for individuals with low digital media literacy, and are likely to be more effective across a broader spectrum of the population than debunking or fact-checking.
- For all its potential for misuse, GenAI is not the first technological development to raise concerns about the potential to spread disinformation, electoral integrity, and information integrity. As with other developments before it, the level of trust in democracy, media, and other institutions remain crucial factors that either amplify or hinder the impact of misleading GenAI.

In this report, we reflect on the presence, manner of use, and challenges of countering GenAI during the 2024 EP elections, while also reflecting on the context in which GenAI played a role in these elections, such as public perceptions of GenAI and its potential to manipulate information integrity. Unlike most elections where GenAI has played a role thus far, the 2024 EP elections were among the first where legislation ranging from the DSA Election Integrity Guidelines to the Code of Conduct for the 2024 European Parliamentary Elections regulated the use of GenAI both for VLOPs and political actors, making it a vital event to examine for insights about the future dynamics between GenAI development and related policy.

For all its potential for misuse, GenAI is not the first technological development to raise concerns about the potential to spread disinformation, electoral integrity, and information integrity.

Incidents of GenAI use in the 2024 European Parliament elections, and techniques and motives

To examine the prevalence, techniques, and motivations behind GenAI use in the context of the EP 2024 elections, we relied on input from researchers based in eight EU member states – France, Germany, Hungary, Italy, Poland, Romania, Spain, and Sweden.

We asked the researchers to answer three basic questions: "What was the prevalence of GenAI in the European Parliament (EP) Elections, which actors used it, and for what purpose?" Comprehensive desk research and interviews with experts further corroborated the experts' findings.

Across the EU, evidence from our research, as supported by findings from other studies, shows that easily identifiable GenAI was most frequently used in France.

One of these most interesting cases involved TikTok accounts impersonating three fictitious relatives of National Rally (Rassemblement National) leader Marine Le Pen and her niece, Reconquest (Reconquête) EP candidate Marion Maréchal, that promoted nationalist sentiments and endorsed their parties, while relying on face-swapping GenAI. These accounts gained significant traction, with some videos amassing over 600,000 views before being deleted.

7 Théophane Hartmann, "Viral Deepfake Videos of Le Pen Family Reminder that Content Moderation Is Still Not Up to Par ahead of EU Elections", Euractiv, 16 April 2024.

8 https://faceswap.dev/

9 Miazia Schueler, Salvatore Romano, Natalia Stanusch, Raziye Buse Çetin, Sonia Tabti, Marc Faddoul & Ibis Lilley,"Artificial Elections – Exposing the Use of Generative AI Imagery in the Political Campaigns of the 2024 French Elections", AI Forensics, 4 July 2024.

The anonymous creator described the project as a social experiment that was "not politically motivated" and only sought to highlight the dangers of disinformation and deepfakes[7]. Many open-source tools exist that could have been used to generate these videos, including Faceswap.dev[8].

Despite protest over these accounts, both the National Rally and Reconquest used unlabeled GenAI images extensively. During May and June, the NGO AI Forensics identified 51 instances of GenAI imagery circulating on VLOPs in France[9]. Most of these images were linked to National Rally and Reconquest, where they were used as a key part of their campaign strategy. This trend of utilising GenAI in political campaigns is notable among right-wing groups across several countries, as highlighted by both our research and findings from the Digital Forensics Research Lab[10]. For instance, as shown by research from several civil society organisations and the University of Amsterdam, the National Rally's initiative "L'Europe Sans Eux" (Europe without Them) used GenAI images across all major social media platforms[11]. None of these images were labelled as generated by AI.

Foreign actors also used GenAI in France to influence public opinion. On 14 February, former Russian President and current Deputy Chairman of Russia's Security Council, Dmitry Medvedev, shared a video on X falsely portraying a France24 journalist as claiming that French President Emanuel Macron refused to visit Kyiv due to a planned assassination attempt by authorities in Ukraine. The low quality of the manipulation likely led to the video being shared as a recording of a television screen. The video relied on GenAI to mimic the presenter's voice, but several features made it easy to identify[12]. The video, which could have been created with tools such as ElevenLabs.io[13], was circulated in pro-Russian Telegram groups, after which it was picked up by pro-Russian outlets (including Russian state-funded newspaper Izvestia, which cited an obscure pro-Russian profile on X[14], and French-registered Pravda.fr), gaining significant traction after Medvedev shared it.

French television broadcasts a report on Macron's cancellation of a visit to Ukraine due to an assassination attempt against him by Ukrainian authorities. So, the evil cocaine addicts wanted to pin Emmanuelle to the ground?

Other cases in France included GenAI footage (once again, relying on face-swapping technology) of Macron dancing in a nightclub and cross-dressing, contributing to the spread of disinformation about his personal life and sexual orientation.

On TikTok, the DFRLab identified two uses of GenAI songs in support of the National Rally

The account @rn.musique published a video of a concert overdubbed with an AI-generated song celebrating the party's lead candidate in the EP elections, Jordan Bardella. The longer version was flagged by @rn.musique as "created using AI", while the shorter version was not[15].

10 Valentin Châtelet, "Far-Right Parties Employed Generative AI ahead of European Parliament Elections", Digital Forensics Research Lab, 11 June 2024.

11 Salvatore Romano, Miazia Schueler, Denis Teyssou, Natacha Farina Groux & Sonia Grillot, "Rassemblement National uses Sensationalizing Generative AI Imagery in its EU Electoral Campaign", 8 June 2024.

12 Sophia Khatsenkova, "Did France 24 Air a Segment Claiming Ukraine Ordered Emmanuel Macron's Assassination?", Euronews, 20 February 2024.

13 https://elevenlabs.io/

14 https://x.com/Jose_FERNANDE8/status/1757627445342007533

In Italy, Matteo Salvini and his party, Lega, created 19 different posts, as identified by Alliance4Europe, using GenAI[16].

Lega used these posts as part of their "Più Italia, Meno Europa" (More Italy, less Europe) electoral campaign, starting with a post by Salvini on 22 May. This post attracted extensive attention, and led to a series of images aimed mainly at inciting nationalistic and Eurosceptical views. In the first post, Europe is represented by a pregnant Jesus Christ – alluding to the political issue of surrogacy practices – while Italy is depicted as a traditional and happy family.

Salvini and Susana Ceccardim, another EP candidate for Lega, used other posts to evoke anti-Islamic sentiments, and to fuel fears about the presence and influence of Muslims in Europe.

This content can be found on various Lega accounts on the Facebook, X, and Instagram platforms. Lega could have used any range of GenAI image-generating tools, including Midjourney, Stable Diffusion, and many others, to create these images. Neither Lega nor the platforms labelled these images as generated by AI.

15 Valentin Châtelet, "Far-Right Parties Employed Generative AI ahead of European Parliament Elections", Ibid., note 10.

16 Saman Nazari, Claudia de Sessa, " Salvini's Electoral Campaign Uses Non-Watermarked AI Images", Alliance4Europe, 6 June 6 2024.

I have become a member of the AfD because there are too many wind turbines.

In Germany, GenAI was almost exclusively used by the Alternative für Deutschland (AfD) party, and mainly on Facebook. It strategically employed these images to foster negative sentiments towards migrants, or to evoke a sense of nostalgia for an idealised, ethnically homogeneous Germany, similar in sentiment to the posts shared by Salvini and Lega[17].

The AfD accounts that used GenAI images were largely regional party associations, and several of the images presented non-existent young people explaining their reasoning behind supporting the AfD party – frequently simply echoing familiar AfD talking points.

In a similar vein, the AfD district association in Esslingen shared an image showing a roasting pig and a group of people along with the hashtags "Enjoyment month" and "German barbecue festival" at the beginning of Ramadan. The AfD party's lead candidate, Maximilian Krah, also regularly used GenAI images on his Facebook page.

17 Felix Heusmann, "Wie die AfD mit KI-Bildern Stimmung macht", Kölner Stadt-Anzeiger, 1 April 2024.

Based on these findings, we observed a consistent trend across France, Germany, and Italy, with far-right political actors using GenAI content as an integral part of their campaigns. This content frequently communicates familiar nationalist, anti-Islamic, and conservative talking points, and the parties or the platforms sharing it do not label or watermark it.

This supports prior research showing that, at least since the autumn of 2023, the extreme right Identity and Democracy group of parties in the European Parliament has used GenAI tools to create synthetic images for social media. These images depict scenarios such as migrants invading the EU, farmers with tractors protesting, and EU politicians in unflattering light. Despite signing a voluntary code in April 2024 to label AI-generated content, nearly all posts violated this guideline. The DFRLab found only one instance that was labeled among dozens of such posts[18].

Finally, while generative AI content played a role in Poland's national parliamentary elections, its use was minimal in the European Parliament elections there. The fact-checking group Demagog uncovered a case of a fake image supposedly showing a roadblock during agricultural protests in March[19]. Shared by conservative figures like journalist Jan Pospieszalski and then-incumbent Law and Justice (PiS) MP Joanna Lichocka, the image received significant attention online. Closer inspection, however, revealed distorted elements (people, vehicles, and unproportionally large flags), indicating it was a fake. The original source of the image remains unknown.

18 Valentin Châtelet, "Far-Right Parties Employed Generative AI ahead of European Parliament Elections", Digital Forensics Research Lab, 11 June 2024.

19 "Czy to Zdjęcie z Protestu Rolników? Nie, to Dzieło AI", Demagog, 22 March 2024.

Policy Framework and response

Concerns about the impact of GenAI on democracies and electoral processes have triggered efforts to counter the negative effects of this technology around the world[20].

For the EP elections, these efforts came from four main sources: the DSA, the newly adopted AI Act, Community Guidelines and policies introduced by online platforms, and voluntary commitments made by political parties and AI providers/deployers/platforms.

All of these regulations provide rules on the transparency and detectability of GenAI content. Despite this, our report reveals that, during the EP Elections, both political parties and online platforms failed to properly label such content. While the challenges VLOPs face in detecting GenAI are valid, it is essential that they ensure that at least GenAI content from politicians and political parties (including deepfakes, but not exclusively) is correctly labeled.

There are also regulations regarding how GenAI models operate, including their data input and safety measures to prevent the production of illegal or harmful content. Most of these requirements are outlined in the AI Act, and will only take effect in 2025 or 2026. It is crucial, therefore, for the European Commission and other stakeholders to closely monitor the development and implementation of self-regulatory tools, such as AI Codes of Practice, during this interim period.

The DSA Response: Guidelines on Electoral Integrity and enforcement actions

Two months before the EP elections, the European Commission issued Guidelines on Electoral Integrity[21] (hereinafter, "Guidelines") within the framework of the DSA[22]. The Guidelines, which are not legally binding, identify GenAI as a systemic risk to civic discourse and electoral processes, and proposes several measures to mitigate the potential of these technologies to mislead voters or manipulate elections through "hallucinations" or inauthentic, biased, or misleading synthetic content. The mitigation measures vary depending on whether VLOPs and VLOSEs (very large online search engines) allow users to create (Para. 39) or solely disseminate (Para. 40) GenAI content. Table No. 1 outlines the proposed measures for each case.

20 The Digital Policy Alert, "Regulating Artificial Intelligence".

21 European Commission, "Guidelines for Providers of Very Large Online Platforms and Very Large Online Search Engines on the Mitigation of Systemic Risks for Electoral Processes Pursuant to Article 35(3) of Regulation (EU) 2022/2065", 26 April 2024.

22 Regulation 2022/2065, Digital Services Act, 19 October 2022.

TABLE NO. 1. MITIGATION MEASURES PROPOSED BY DSA GUIDELINES ON ELECTORAL INTEGRITY

Case	Who does this apply to?	Proposed mitigation measures
VLOPs and VLOSEs whose services can be used for the creation of deceptive, biased, false, or misleading GenAI content (Para. 39)	Some examples: ● Google's Gemini[23] ● Microsoft Bing's AI Assistant Copilot[24], and image generator Image Creator, by Designer[25] ● TikTok's Symphony Assistant[26] ● YouTube's AI creator tools[27]	● **Ensure GenAI content is detectable** using techniques like watermarks, metadata, or cryptographic methods, especially for content related to candidates, politicians, or political parties ● **Base AI-generated information on reliable sources,** and provide links to official electoral authorities to minimise "hallucinations" ● **Encourage users to verify electoral information** with authoritative sources ● **Conduct and document red-teaming exercises,** focusing on electoral processes ● **Monitor GenAI systems for safety and factual accuracy,** particularly concerning electoral content ● **Implement safeguards,** such as prompt classifiers, content moderation, and other filters, to prevent the misuse of GenAI for creating illegal or manipulative disinformation during elections ● For text-based GenAI content, **include links to sources**, so users can verify the information's reliability and context.
VLOPs and VLOSEs whose services can be used to **disseminate** deceptive, false, or misleading GenAI content (Para. 40) VLOPs and VLOSEs whose services can be used to disseminate deceptive, false, or misleading GenAI content (Para. 40)	All VLOPs and VLOSEs, particularly social media platforms (X, TikTok, Facebook, Instagram, LinkedIn, YouTube, etc.	● **Adapt and enforce terms and conditions** to significantly reduce the reach and impact of GenAI content that spreads disinformation or misinformation about electoral processes. This includes publicly disclosing measures such as labeling, marking, demoting, or removing content. ● **Clearly label deepfakes,** and provide users with an easy way to identify GenAI labels. These labels should persist even after the content is reshared. ● **Update advertising systems** to allow advertisers to clearly label GenAI content or require such labels for ads ● **Adapt content moderation processes and algorithmic systems to detect AI-generated or manipulated content,** using techniques like watermarks, metadata, cryptographic methods, logging, and fingerprints ● Implement **media literacy measures.**

23 Gemini Models, Gemini.
24 Microsoft, Copilot.
25 Microsoft, Imager Creator from Designer.
26 TikTok, TikTok Symphony.
27 Toni Reid, "Made On YouTube: Empowering anyone to Create on YouTube", 21 September 2023.

In addition to issuing the Guidelines, the European Commission sent requests for information to six VLOPs and two VLOSEs to gather details on their GenAI policies[28].

It also held an election readiness round table to test VLOPs and VLOSEs incident response mechanisms to GenAI threats[29]. So far, the Commission has not taken further action on these cases.

Despite these efforts, the DSA framework has notable limitations in addressing the risks associated with GenAI. While the Guidelines assist the Commission in assessing compliance and provide companies with a framework for their efforts, they remain soft law, and are not legally binding. Furthermore, the Guidelines primarily focus on VLOPs and VLOSEs, leaving out other online platforms, such as Telegram, that are not classified as such, but can still play a significant role in spreading GenAI content. The Guidelines only suggest that these platforms might consider using the recommended measures as "inspiration"[30].

Moreover, popular GenAI systems, such as Midjourney, Dall-E, or Stable Diffusion, are not covered by the DSA, as they do not qualify as intermediary services. There is ongoing debate among experts[31] as to whether large language models should be covered by the DSA, particularly when they are analogous to search engines, such as the recently launched SearchGTP[32]. This interpretation remains contested, however.

The AI Act aims to address this loophole and manage the risks associated with GenAI from the ground up, focusing on the models and systems that generate the content. Regulations governing GenAI technologies, however, will not come into effect until 2025 and 2026.

AI ACT: RULES FOR CHATGTP AND FRIENDS

The AI Act[33] introduces hard-law rules for GenAI models and systems. These regulations align closely with some of the measures discussed earlier, as the Commission integrated certain obligations from the AI Act directly into the Guidelines, to ensure consistency.

On the one hand, the Act sets transparency obligations. Chatbot providers must inform users they are interacting with AI, and not a human, while providers of synthetic content must label outputs in a machine-readable format that indicates artificial creation or manipulation[34]. Since these requirements will not come into force until two years after the AI Act takes effect (i.e., August 2026), the AI Act encourages the AI Office to draft EU-level Codes of Practice for detecting and labeling such content during the interim[35].

The Act also imposes additional rules on general-purpose AI models (GPAI)[36], which form the base of most GenAI systems. GPAI providers must prepare technical documentation, share information with downstream users (such as deployers), and comply with copyright laws. Recognising that some GPAI models might pose systemic risks, such as negatively impacting democratic processes[37], the Act introduces a "systemic risk" category for models with significant "high-impact capabilities" or substantial market influence in the EU[38]. For these models, the Act mandates regular risk assessments, proactive risk mitigation, continuous incident monitoring, and robust cybersecurity practices. Obligations for GPAI will not be applicable until August 2025.

Notably, open-source GPAI models – those publicly available under free or open licenses, with accessible architectures – are relatively underregulated. They are only required to meet obligations if they pose a systemic risk or fall into the High-Risk AI category.

28 European Commission, "Commission Sends Requests for Information on Generative AI Risks to 6 Very Large Online Platforms and 2 Very Large Online Search Engines under the Digital Services Act", 14 March 2024.

29 European Board for Digital Services, "Report on the European Elections. Digital Services Act and Code of Practice on Disinformation", p. 11, July 2024.

30 European Commission, Guidelines on Electoral Integrity, paragraph 17.

31 Beatriz Botero Arcila, "Is it a Platform? Is it a Search Engine? It's Chat GPT! The European Liability Regime for Large Language Models", Journal of Free Speech Law, Vol. 3, Issue 2, 2023.

32 OpenAI, SearchGPT Prototype, 25 July 2024.

33 Regulation EU 2024/1689, Artificial Intelligence Act, 13 June 2024.

ONLINE PLATFORMS' RESPONSES: GenAI POLICIES

Online platforms have taken steps to align with the European Commission Guidelines on Election Integrity by updating their internal policies and community guidelines. Annex 1 provides a detailed overview of the policies adopted by five major VLOPs to address the risks associated with the dissemination and, to a lesser extent, the creation of GenAI content during the 2024 EP Elections. Some key trends included:

- Community guidelines and policies apply to all content. All platforms assert that they will take action against any content that violates their policies (including cases of disinformation or hate speech), whether AI-generated or not.
- Bans on deceptive content. All platforms ban realistic content (e.g., deepfakes) created with AI or otherwise, if it intends to confuse or deceive users and potentially cause harm.
- Labeling requirements. While non-deceptive AI-generated content is allowed, most platforms require that realistic images, video, or audio (i.e., deepfakes) be labeled as such. Notably, X will only label synthetic content if it deceives users or is potentially harmful, but does not pose a serious enough risk to require removal.
- AI-generated ads. Meta and YouTube require advertisers to disclose cases where their ads were generated or altered using AI.
- User labelling features. All platforms, with the exception of X, provide users with tools to label AI-generated content. Typically, they mandate the labeling of realistic images, video, or audio (e.g., deepfakes), while labeling other types of AI-generated content is only encouraged.

While this report cannot conclusively determine whether the VLOPs' policies fully meet the mitigation measures recommended by the DSA Guidelines, it is important to highlight that most of the GenAI cases included in this report went unlabeled.

Some platforms have acknowledged this shortfall, arguing that automatic detection of GenAI content remains a significant challenge, due to the lack of adequate technologies. Platforms can automatically label content generated by their own AI systems (Meta and TikTok are already doing it), but identifying and labeling content produced by other GenAI platforms is considerably more difficult[39].

34 AI Act, Article 50 (1) and (2).

35 AI Act, Article 50 (7).

36 AI Act, Article 52.

37 AI Act, Recital 110.

38 AI Act, Annex XIII. Criteria for the Designation of General-Purpose AI Models with Systemic Risk Referred to in Article 51.

In February, Nick Clegg, Meta's president of global affairs, said "it's not yet possible to identify all AI-generated content"[40]. While cross-industry initiatives have improved the identification of AI-generated images from platforms such as Google, OpenAI, Microsoft, Adobe, Midjourney, and Shutterstock, GenAI tools for audio and video have not incorporated detection signals as effectively. This makes labeling more challenging. Also, users can employ techniques to remove invisible markers from content, further complicating detection efforts.

As explored in section VI, detection challenges are real. In the context of elections, however, platforms should at least ensure that synthetic content, including AI-generated materials shared by politicians or political parties, is appropriately labeled, regardless of whether it is in the form of deepfakes. This is a fundamental rule of transparency and can be reasonably achieved using both automatic and manual methods, as demonstrated by the work of some civil society organisations.

VOLUNTARY COMMITMENTS BY POLITICAL PARTIES, AI PROVIDERS, AND ONLINE PLATFORMS

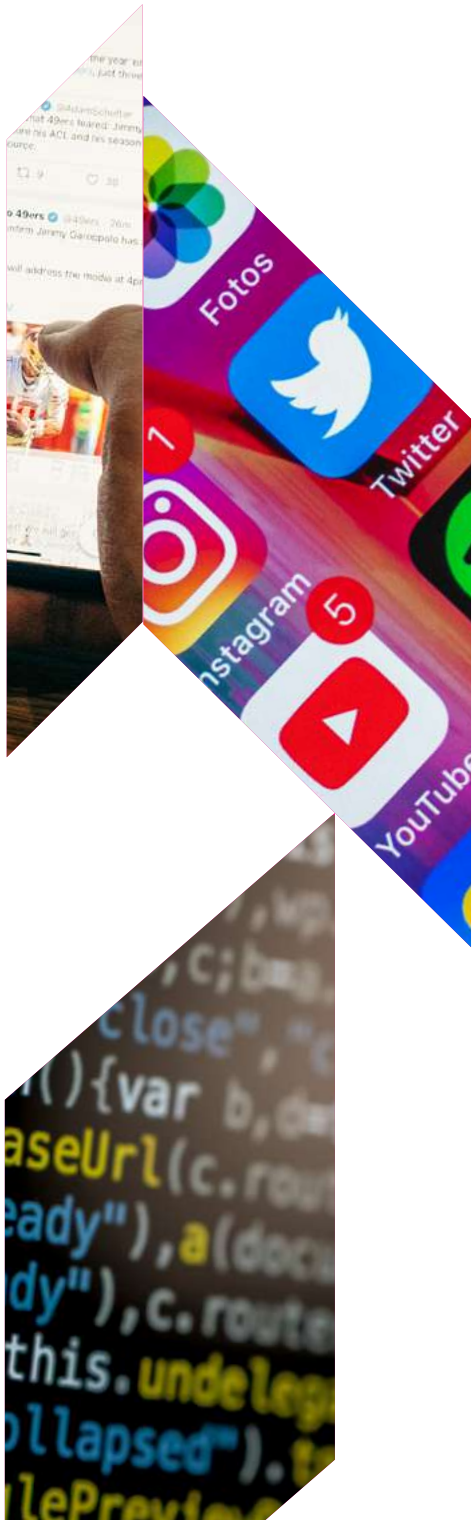
The hype around GenAI risks during the so-called "election year" prompted the adoption of self-regulatory measures.

On 16 February, at the Munich Security Conference, 25 tech companies, including Meta, OpenAI, Microsoft, and X, signed the Tech Accord to Combat Deceptive Use of AI in the 2024 Elections[41]. The instrument includes measures for collaborating on tools to detect and address deceptive GenAI content, driving educational campaigns, and increasing transparency.

In parallel, in April, political parties participating in the EP elections signed a Code of Conduct developed by International IDEA[42]. The code aimed to ensure the integrity and fairness of the EP election campaign by requiring, among other measures, that GenAI content be clearly labelled, with watermarks and signals identifying their origin and/or creator strongly encouraged.

Despite these commitments, far-right political parties across France, Germany, and Italy not only used GenAI content as an integral part of their campaigns, but largely also did not label it.

More recently, on 30 July, the European AI Office announced it would begin the process to draft the General-Purpose AI Code of Practice[43]. This code will align with the AI Act's obligations, helping GenAI providers demonstrate compliance when their obligations come into effect in 2025.



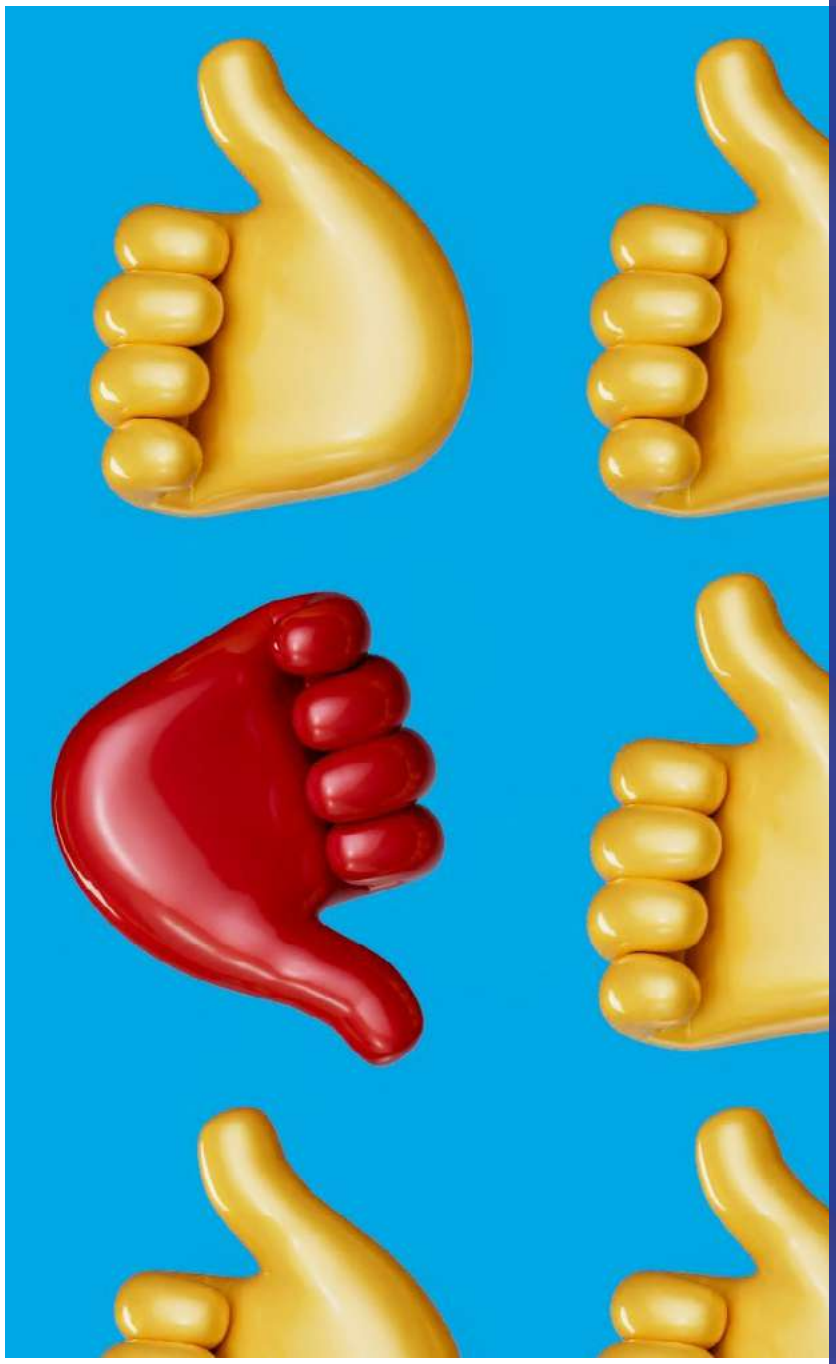
39 Eliza Gkritsi, "The Brief – DSA is Branching into GenAI Regulation for EU Elections", Euractiv, 30 May 2024.

40 Nick Clegg, Labeling AI-Generated Images on Facebook, Instagram and Threads, Meta, 6 February 2024.

41 Munich Security Council, "A Tech Accord to Combat Deceptive Use of AI in 2024 Elections", February 2024.

42 International IDEA, in collaboration with the European Commission, Code of Conduct for the 2024 European Parliament Elections, 9 April 2024.

43 European Commission, AI Act: Participate in the Drawing-Up of the First General-Purpose AI Code of Practice, 30 July 2024.



Public perception of GenAI impact on information integrity online

In assessing the prevalence and technological challenges of countering unlabeled GenAI, we should also consider public perceptions about the potential impact of GenAI on information integrity, and how citizens assess their own resilience to misleading GenAI content.

In a study conducted in Germany by Syzygy Group on GenAI perceptions and knowledge, only eight per cent of respondents could identify a realistic GenAI picture of a non-existent individual. Perhaps significantly more concerning is the fact that the same study found that only two-thirds of Germans were aware of the concept of GenAI, with a similar proportion reporting they had heard of ChatGPT. It is not surprising, therefore, that nine in ten Germans believe they should have the right to know whether AI has generated or altered the content they are exposed to. Additionally, three-quarters of Germans support a 'Blade Runner'-style law, which would make it illegal for AI to conceal its identity and impersonate a real human[44].

A recent Ofcom survey of 3,000 UK residents eight years of age or older revealed similar trends to those in Germany. Fewer than one in ten people 16 years of age and older say they are confident in their ability to identify a deepfake, although younger children, ages 8-15, are more likely to express confidence (20 percent)[45].

Looking beyond Europe, a survey of 1,020 U.S. citizens showed that 78 percent expect to see GenAI abuse in the 2024 U.S. Presidential elections, and 69 per cent said they are not confident that most voters can detect such content, yet 42 per cent of those surveyed stated that they are confident in their own ability to detect GenAI[46].

These studies, along with the challenges that trained practitioners, experts, and platforms face in identifying GenAI, strongly underscore the need for media literacy initiatives across the public spectrum, and especially among individuals with already low levels of digital media literacy, including elderly voters.

44 Dr. Paul Marsden,"A First Survey of Public Perceptions of GenAI in Germany", Syzygy Group, 23 March 2023.
45 "A Deep Dive into Deepfakes that Demean, Defraud and Disinform", Ofcom, 23 July 2024.
46 " AI & Politics '24", Elon University, May 15 2024.

What are the risk factors of GenAI for tipping an election?

GenAI is one of many technological developments that raise concerns about the possibility of disinformation and the manipulation of public opinion at scale, especially during key societal moments, such as elections.

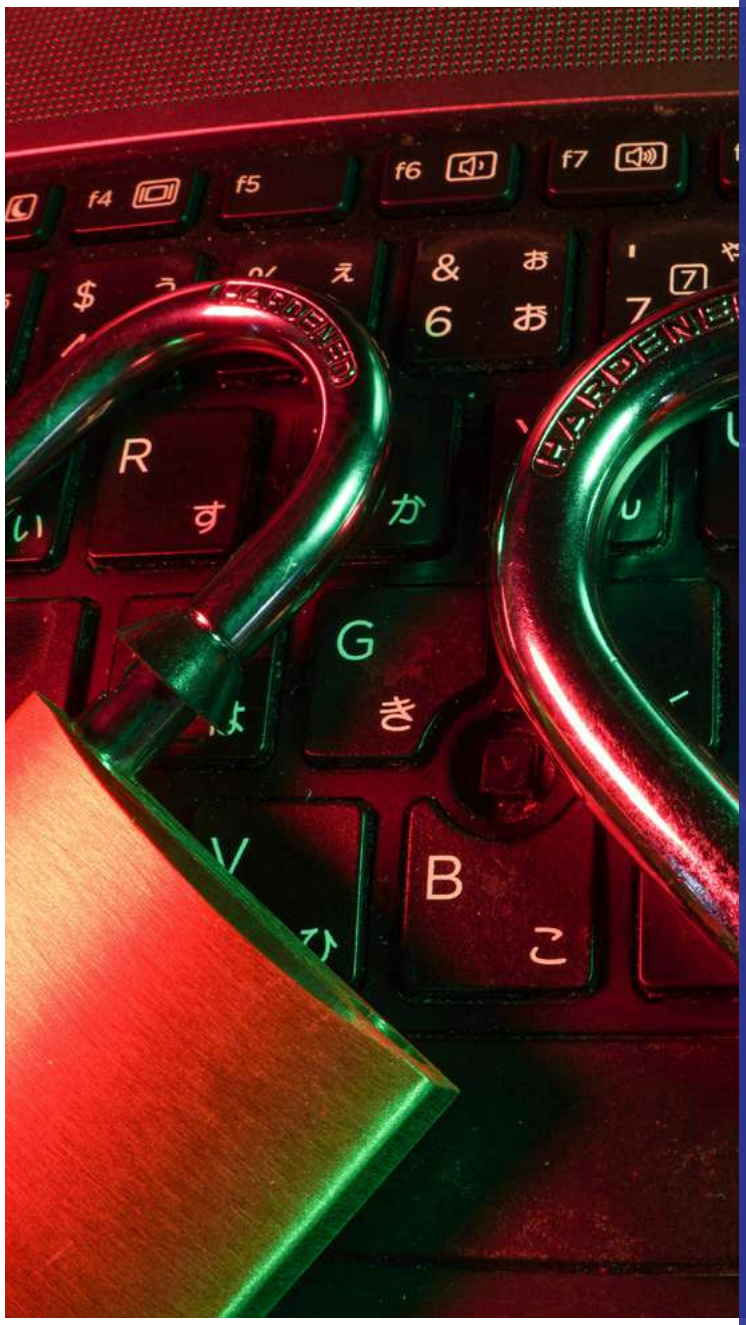
In an interview, Dr. Thorsten Quandt, Professor of Communication Studies at the University of Münster, noted that, like other technological developments (such as low-intelligence bots) before it, GenAI will have an impact. Still, he said, it will not necessarily be a "game-changer" as many anticipate. He argued that trust in democracy, the political system, and traditional journalism are highly relevant factors in determining a society's susceptibility to any type of disinformation, including that supported by GenAI. This is a primary reason why many populist political actors work globally to undermine trust in the media.

Quandt further cited research showing that pre-bunking works better than debunking as a tool to improve media literacy. Debunking and fact-checking often lead to cognitive dissonance, affecting only those "in doubt", and do not impact fervent supporters of populist political actors, who generally have low institutional trust across the board. Quandt ended our interview on a positive note. He claimed that, while the expected scalability and consequent presence of GenAI content may lead to an overall deterioration of trust in institutions, it may also inadvertently increase trust in "good news sources" that consistently prove their "trustworthiness". He cited research from the University of Mainz showing that trust in traditional journalism in Germany has increased across almost all parts of society in recent years.

Recent research shows that threats to the way we understand and process information can't be looked at separately without considering what they reveal about the overall quality of our news environments[47]. The study shows that vulnerability to misleading information aligns "remarkably well" with the different ways media systems are structured around the globe. Notably, northern European countries "exhibit greater epistemic resilience", here understood as resilience to misleading information, "while the US, Spain, and Eastern Europe face more vulnerability". The study also provides strong evidence that ideological polarisation and high degrees of connection between political and media systems increase levels of susceptibility to mis/disinformation.

Thus, while GenAI is likely to play an increasingly important role in future elections and other key societal moments, we can best understand its ability to affect societal developments by taking a broader view of the media and political environment in which it is disseminated.

47 Julien Labarre," Epistemic Vulnerability: Theory and Measurement at the System Level", Political Communication, June 2024.



Methods and challenges in identifying GenAI content

Unlike the relatively easily identifiable GenAI campaign materials mentioned above, manipulative actors can use GenAI to create more convincing content.

Even now, manual content reviews are the most reliable methods for identifying GenAI imagery, video, voice cloning, and other instances of such manipulation. Existing technical tools, however, including VeraAI[48], inVID[49], Forensically[50], TinEye[51], TrueMedia.org[52], and Google reverse image search, as well metadata inspection, can support these efforts.

For instance, AI Forensics used a combination of manual reviews from three reviewers – inVID, Google reverse image analysis, and TrueMedia.org – showcasing the cumbersome process necessary for moderation with the technology currently available to civil society. In an interview, Salvatore Romano, Head of Research at AI Forensics, explained that even processes based on cross-referencing, such as this one, still do not provide a "final answer", showcasing the challenges civil society organisations and platforms alike face in detecting the use of GenAI.

Other approaches for detecting GenAI involve inspecting the metadata of content (the hidden layer of information that accompanies every media file we encounter), which provides information on the origin of the file, its size and format, and other distinctive properties. For instance, checking timestamps in the metadata can be crucial. If a timestamp does not match the actual event depicted in the photo, it may reveal that the image was not taken at the scene. Moreover, metadata analysis is useful because it is scalable. If many files were created at the same time, this could indicate they were generated using AI, since human content creation is more sporadic and random. Additionally, authentic content often has extensive metadata, including location data, equipment used, settings, and more. As with other methods, however, metadata analysis is not a silver bullet. Skilled manipulators can alter or strip away this information, making the detection process trickier, and most social media platforms also remove this data. Both DRI and AI Forensics have published useful guidebooks for GenAI detection for civil society organisations and other practitioners[53].

48 https://www.veraai.eu/home

49 https://www.invid-project.eu/

50 https://29a.ch/photo-forensics/#forensic-magnifier

51 Thttps://tineye.com/

52 https://www.truemedia.org/

53 "A Manual Guide to Detecting Generative AI Imagery", AI Forensics; Jan Nicola Beyer, Beatriz Almeida Saab, Lena-Maria Böswald, " Synthetic Media Exposed: A Comprehensive Guide to AI Disinformation Detection", October 2023.

ANNEX 1: EUROPEAN PARLIAMENT ELECTIONS GENERATIVE AI POLICIES IN FIVE VERY LARGE ONLINE PLATFORMS (VLOPS)

Updated 8.8.2024

	Meta	TikTok	YouTube	X (Formerly Twitter)
Additional content moderation rules for GenAI content?	**Partially.** AI-generated content is eligible for review by independent fact-checking organisations. If deemed "faked, manipulated, or transformed" audio, video, or photos, Meta will label and down-rank it in the feed. There are no additional rules in the Manipulated Media policy specifically for AI-generated content. Existing rules apply to all content (GenAI or not) and prohibit: ● videos depicting a person saying words they did not actually say; and ● videos that combine, replace, or superimpose content onto another video to make it appear authentic.	**Yes.** TikTok Community Guidelines do not allow the dissemination of certain GenAI content (labelled or not), including: ● photorealistic content depicting people under 18 years of age; ● photorealistic content depicting adults without their permission; and ● content that falsely portrays authoritative sources, crisis events, or public figures in misleading contexts, such as being bullied, making endorsements, or being endorsed.	**No.** YouTube relies on its long-standing policies that prohibit technically manipulated content that misleads viewers and poses a serious risk of harm.	**No.** According to X's Synthetic and manipulated media policy, sharing synthetic, manipulated, or out-of-context media that may deceive or confuse people and cause harm is prohibited. Media that is not removed may be labelled to provide authenticity and context. To determine whether there are violations, X analyses whether the content: ● is significantly and deceptively altered, manipulated, or fabricated; ● is shared in a deceptive manner or in a false context; or ● is likely to cause widespread confusion on public issues, to impact public safety, or to cause serious harm. Non-violations: Memes, satire, animations, illustrations, cartoons, commentary, reviews, opinions, and counterspeech, provided they do not cause significant confusion about authenticity.
Detection and labelling of content generated with their own GenAI features	**Yes.** Photorealistic images created with Meta AI will be labelled with visible markers, invisible watermarks, and IPTC metadata. Audio and video not mentioned.	**Yes.** TikTok labels AI-generated content made with TikTok AI effects. They also renamed TikTok AI effects to explicitly include AI in their name.	**Partially.** YouTube states that it may proactively add a label if a creator does not disclose the use of GenAI, especially when altered or synthetic content could confuse or mislead viewers.	**Not applicable.** X's AI feature is the chatbot Grok.

ANNEX 1: EUROPEAN PARLIAMENT ELECTIONS GENERATIVE AI POLICIES IN FIVE VERY LARGE ONLINE PLATFORMS (VLOPS)

Updated 8.8.2024

	Meta	TikTok	YouTube	X (Formerly Twitter)
Detection and labelling of content generated with other GenAI platforms	Currently collaborating with industry partners to develop common technical standards for detecting and labelling photorealistic AI-generated images from other GenAI platforms. AI-generated audio and video are not mentioned. AI-generated or altered images, video, or audio that pose a high risk of deceiving the public may have more prominent labels.	In May, TikTok announced it was "starting to automatically label AI-generated content when upload from certain other platforms". To do so, they partnered with the Coalition for Content Provenance and Authenticity (C2PA).	It also mentions that it collaborates across the industry – C2PA, to help increase transparency around digital content.	**Not mentioned.** X only states it may label synthetic, manipulated, or out-of-context media that may deceive or confuse people and cause harm.
User Labeling Feature?	**Yes.** Users have a feature to label GenAI video or audio. Meta requires users to label content they share that has photorealistic video or realistic-sounding sound, and that has been digitally generated or altered, including with AI. Failure to do so may result in penalties. Users are not, however, required to label AI-generated images. Meta states that its systems may automatically detect and label such content.	**Yes.** TikTok provides users with an AIGC label. A label is mandatory only for AI-generated content containing realistic images, audio, and video. For other types of AI-generated content, labelling is encouraged, but not required.	**Yes.** YouTube introduced a tool in Creator Studio for users to disclose content made with altered or synthetic media, including GenAI. Disclosure is mandatory for realistic content (deepfakes). It is not required for unrealistic, animated, special effects, or GenAI-assisted content. Following the disclosure, YouTube adds a transparency label. For elections content, the label will appear on the video itself and in the video description.	**No.**
Measures for GenAI political ads?	**Yes.** Meta does not allow advertisers to use its own GenAI features for ads related to elections or politics. Advertisers must disclose the – use of third-party GenAI systems to create photorealistic images, videos, or realistic-sounding audio for election or political ads. Debunked GenAI ads are not allowed.	**Not applicable.** TikTok does not allow paid political advertising, and accounts belonging to politicians or political parties are not allowed to advertise.	**Yes.** Advertisers must disclose when their election ads include digitally altered or generated materials.	**No.**