BRIEFING

European Parliament

Children and deepfakes

SUMMARY

Deepfakes – videos, images and audio created using artificial intelligence (AI) to realistically simulate or fabricate content – are booming on the internet. They are becoming increasingly accessible, as what previously required powerful tools can now be done with free mobile apps and limited digital skills. At the same time, they are becoming increasingly sophisticated and therefore more difficult to detect, especially audio deepfakes. While deepfakes have applications in entertainment and creativity, their potential for spreading fake news, creating non-consensual content and undermining trust in digital media is problematic, as they are evolving faster than existing legislative frameworks. A projected 8 million deepfakes will be shared in 2025, up from 500 000 in 2023. The European Commission states that pornographic material accounts for about 98 % of deepfakes.

Deepfakes pose greater risks for children than adults, as children's cognitive abilities are still developing and children have more difficulty identifying deepfakes. Children are also more susceptible to harmful online practices including grooming, cyberbullying and child sexual abuse material. This highlights the need for legal action and cooperation, including developing the tools and methods needed to tackle these threats at the required scale and pace. Furthermore, there is a growing need for enhanced generative AI literacy for children, educators and parents. There is also a need for increased industry efforts and better implementation of relevant European Union (EU) legislation such as the Artificial Intelligence Act and the Digital Services Act. Monitoring indicators on children's online use at the EU level are currently non-existent, highlighting the need for their implementation.

IN THIS BRIEFING

EPRS | European Parliamentary Research Service

Author: Mar Negreiro
Members' Research Service

PE 775.855 – July 2025

EN

Evolution of deepfake technology

Deepfake technology creating media content – such as images, audio, or video – that has been generated or manipulated using artificial intelligence (AI), often to appear authentic, utilises advanced generative AI to create hyper-realistic synthetic media resulting in audio or visual content that poses significant challenges across various sectors. Deepfakes simulate and distort reality, often appearing authentic.

This advancement is largely powered by cutting-edge AI technologies, including autoencoders and generative adversarial networks (GANs). Content created using a GAN – a type of AI that learns to produce highly realistic images, video, or audio by training on large datasets – is the main technology behind many deepfakes. These deep learning generative models have progressed to a stage where distinguishing fake images and videos from genuine ones has become difficult, posing many risks for society at large.

Even though deepfake technology started developing in academic institutions over two decades ago, it was not until 2017 that the term 'deepfake' gained popularity, thanks to a Reddit user who started a channel with that name. The channel featured videos created by its members using celebrities as the basis for pornographic content, leading to the group being banned. Broadly speaking, deepfakes can be categorised into two types: those that transform existing media and those that are generated entirely from scratch by AI models, such as image-generation systems, without relying on a pre-existing source.

GANs involve two algorithms working together. These include two competing neural networks: a generator that creates fake images and a discriminator that tries to identify real from fake. This forces continuous improvement, as the generator constantly evolves its output to outwit an improving discriminator. Initially, deepfakes were easily identifiable, often characterised by manipulations of facial expressions and lip movements. However, with rapid advancements in AI models, deepfake technology has evolved exponentially, creating realistic and indistinguishable synthetic digital content.

Another technique often used in addition to GANs, especially for tasks like face-swapping, is the autoencoder – a neural network architecture consisting of an encoder and decoder. When creating a deepfake, the system first trains on thousands of images of two different faces: the source and the target. The encoder part of the network learns to reduce these facial images to their essential features and distinctive traits. It then extracts shared features and swaps facial details while preserving realistic movement and expressions. The decoder creates convincing but distinguishable deepfakes, not as realistic as GANs.

Additionally, generating deepfake synthetic media involves creating both visual and audio content. There are different types of generation processes, including algorithms that leverage sophisticated AI techniques.[1]

With the rise of AI tools and software platforms – ranging from open-source programmes to proprietary systems and mobile applications – deepfake creation has become increasingly accessible and user-friendly, requiring little more than a mobile device and the right software, and catering to both novice and professional users. Most of the main deepfake platforms state that their service is not intended for children under 18 and that minors should use them under parental supervision. This raises concerns about their privacy.

Spread and societal impact of deepfakes

In recent years, deepfakes have doubled in number every six months. A projected 8 million deepfakes will be shared in 2025, up from 500 000 in 2023. Europol estimated that 90 % of online content may be generated synthetically by 2026 as deepfakes spread rapidly through social media platforms, messaging apps and video-sharing platforms, blurring the line between reality and fiction. This makes them a powerful tool for spreading misinformation and disinformation. Additionally, recent studies have found that, on average, people struggle to distinguish between synthetic and authentic media, with the mean detection performance close to a chance level of 50 %. Accuracy rates worsen when the stimuli contain any degree of synthetic content, feature foreign languages, or when the media type is a single modality (visual, audio or text).

According to a survey, even IT experts incorrectly perceived 62 % of AI-generated content as being created by humans.

In the United States (US), a survey has shown that 70 % of teenagers have used generative AI tools. Similarly in the United Kingdom (UK), a large majority of teenagers (four out of five) have also used generative AI tools. Over 50 % of surveyed teens used AI text generators and chatbots, 34 % used AI image generators, and 22 % used video generators. Younger users are more intense users than adults. While only 9 % of adults feel confident in their ability to identify deepfakes, a slightly higher percentage (20 %) of children aged 8-15 report similar confidence. According to Ofcom research, 50 % of UK children aged 8-15 report having seen at least one deepfake in the last six months. There is also a high and ever-increasing demand for sexually explicit deepfakes. When it comes to synthetic sexual content, one in seven users who saw synthetic content saw sexual deepfakes, mainly featuring women, and 17 % of the synthetic sexual content involved a minor. Moreover, only 37 % of parents whose children use AI tools are aware of it, and nearly 25 % of parents incorrectly believed their children were not using AI tools. Most parents have not discussed the use of AI tools with their children.

Children are particularly vulnerable to synthetic content, such as deepfakes, which can make them more exposed to harmful online practices like grooming and cyberbullying, as well as to child sexual abuse material (CSAM). However, at the EU level, there have been no surveys conducted to date on the use of deepfakes by adults, children or teenagers. Only some anecdotal surveys at the national level are currently available on broad topics like AI and the future of work.

Malicious uses of deepfakes

While deepfakes have legitimate applications in entertainment and other areas, their potential for spreading fake news, creating non-consensual content, committing cybercrimes and undermining trust in digital media can turn them into tools for dangerous societal misuse. Misuse of deepfake technology includes financial crimes, extortion, harassment and the creation of pornographic deepfakes. Moreover, deepfakes pose particular risks for children, whose cognitive capacities are still developing and who have difficulty identifying deepfakes as they are become increasingly sophisticated. This makes it challenging for children to distinguish human-generated content from AI-generated (synthetic) content. The EU strategy Better internet for kids highlights one misuse of deepfakes among children as spreading rumours, including creating content to falsely associate someone with an act they did not commit, as part of a harassment campaign or cyberbullying. A recent report warned of an amplification of different types of misuses involving deepfakes. Below are outlined some of the main threats that can be exacerbated.

Fraud and manipulation

The evolution of deepfake technology, particularly with the rise of generative AI, has significantly reshaped the landscape of digital media manipulation and cybercrime, including deepfake scams. Generative AI is making it easier for fraudsters to create deepfakes, solicit money and perpetrate other types of fraud. Most recently, diffusion models in AI platforms have emerged as widely available tools for threat players to conduct deepfake campaigns. Researchers expect increasing use of generative AI to create deepfakes in the coming years. Deepfakes may enhance identity theft, phishing, investment and recruitment scams. According to a report, a deepfake attack occurred every five minutes in 2024.

Deepfake-based fraud is affecting businesses all over the world. In 2024, 49 % of companies experienced audio and video deepfakes. New research shows that, in 2024, the use of generative AI-based deepfakes increased by 118 % and that 90 % of US companies experienced cyber fraud. Other research shows that deepfake fraud increased by 1 740 % in the US in 2022.

The most convincing and hardest to detect AI-generated content are voice deepfakes, allowing cybercriminals to capture voice samples from interviews, podcasts or social media to replicate voices very realistically. They easily obtain voices, since 53 % of people share their voices online or through recorded notes at least once a week. Feeding these recordings into neural network-based models, they can generate AI voices closely resembling the original speaker's voice, enabling voice scams and voice phishing (i.e. vishing), where criminals request urgent financial transfers or sensitive internal data while pretending to be someone else, including children and other family members. As voice deepfakes become more accurate, targeted vishing attacks increase. With real-time impersonation of family members, employees or customer service representatives, the fraud can lead to unauthorised transactions, data breaches and social engineering schemes. A survey found that one in four adults had experienced or known someone affected by an AI voice cloning scam, with 70 % unsure of their ability to distinguish cloned voices.

Likewise, deepfakes are being used to create fake video calls that could mislead employees. In one case, it resulted in a company losing over US$25 million.

This is also happening in electoral campaigns and elections around the world as part of broader online information manipulation. Children may be misled by manipulated videos or images exposing them to fake news and disinformation, thereby confusing their understanding of reality and affecting their perception the world.

Data protection and privacy

The EU has championed the protection of citizens' privacy since the entry into force of the General Data Protection Regulation (GDPR). When considering children's data protection and privacy, it is important to note that children may not comprehend the concept of privacy. Therefore, it is likely that they may disclose personal information when they interact with AI, without there being any specific measures in place to protect them against it. According to a report, 67 % of respondents agree that students might share personal information with generative AI, and several national data protection authorities have started taking action concerning these data collection practices.

Some experts suggest that parents educate their children about the risks of deepfakes and ensure they implement strong privacy settings on their social media accounts. For instance, personal photos or videos shared online can be exploited to create deepfakes, putting children at risk of being impersonated or targeted. If this were to happen, it could potentially give criminals access to children's personal accounts and allow them to perpetrate fraud or manipulate family members through vishing attacks, among other things.

Regarding measures to ensure the protection of minors from age-inappropriate content, the Italian data protection authority imposed a temporary restriction on OpenAI in an attempt to safeguard its users, especially minors, when personal data was processed.

Cyberbullying, grooming and increased CSAM online

Children face various types of harm from deepfakes, which can result in social collusion, bullying or harassment. Experts warn that the mental health impact of deepfakes can be just as severe as that of real content, causing anxiety, depression and sometimes even post-traumatic stress disorder.

Deepfakes enable new forms of cyberbullying, as they can be used to create embarrassing or humiliating content about a child, leading to online harassment. This issue is further complicated by the fact that children are often both victims and perpetrators of cyberbullying. According to a report, 62 % of respondents were worried about the potential for generative AI to be used for cyberbullying. Encountering a deepfake of themselves or someone they trust can be deeply distressing and cause emotional harm to children. According to a survey by the Center for Democracy and Technology, 40 % of US students and 29 % of teachers reported being aware of deepfakes depicting people they knew that were shared in the last school year. Moreover, 15 % of students and 11 % of teachers were aware of intimate or sexually explicit deepfakes.

Women and girls are particularly vulnerable to deepfakes, especially those of a pornographic nature. This includes the 'nudify' apps or websites used by minors, often at schools, to create and share naked pornographic pictures of their classmates from previously taken social media images. According to a company that analyses social networks, millions of users visit more than 100 'nudify' sites online each month, making them a major driver of the deepfake economy. AI can generate highly realistic nude images, leading to cases of sexual harassment and sextortion that are also alarmingly increasing online. In these cases, victims are coerced into paying money to prevent the distribution of deepfakes about them. Cryptocurrencies are facilitating these transactions, making it easier for developers and users to operate anonymously. Minors, in particular, tend to yield to sextortion due to fear of their parents' reaction and whether they will believe in the falseness of the deepfakes. International law enforcement agencies, such as Interpol, the FBI and Europol, have issued warnings regarding this alarming trend.

Generative AI is also being used to create CSAM, generating hyper-realistic deepfake images and videos from existing content or from scratch. The increase in AI-generated online child sexual abuse is already a growing challenge for law enforcement. Europol has warned about the rise in AI-generated CSAM. The Internet Watch Foundation's report from July 2024 shows that 90 % of images assessed by analysts were realistic enough to be considered under the same law as real CSAM and contained more images in the most severe category than in 2023.[2] It has identified open-source AI models as the tool of choice for paedophiles to create CSAM. In the UK, among surveyed adults who reported being exposed to sexual deepfakes in the last six months, 17 % thought they had seen images portraying minors.

According to researchers, CSAM-based deepfakes might be as harmful as real CSAM for the victims' mental health, similar to how cyberbullying affects individuals. They can also be totally new or synthetic, posing challenges for legal systems and law enforcement agencies' resources and investigations. Moreover, deepfakes can also be used for grooming, as offenders use them to approach children online, pretending to be other children or individuals, and coercing them into producing sexual material.

Mitigating risks associated with deepfakes

Policymakers working to mitigate harm to individuals from AI-generated fake content are faced with several key challenges. Below are some of the efforts to counter these challenges through technology, legislation and education.

Technology

Detecting deepfakes: The race between deepfake generation and detection

Generally, there are two approaches to combating harm inflicted by deepfakes: detection and prevention. Within prevention, there are transparency techniques that seek to achieve transparency through means such as watermarking and the building of blockchain frameworks. Seen by many as a core part of the solution, watermarking synthetic content involves embedding a unique signal to identify that content has been AI-generated, letting people know who owns and created the content. Additionally, systematically tagging whether images are generated by AI or a human could help researchers and platforms better understand the prevalence of synthetic contents on the internet. The main downside of watermarks on AI-generated content is that they could be easily removed. Companies such as Google are developing watermarks that are invisible to the human eye and thus more difficult to remove. The problem is that, for this solution to work, it has to be consistent across all industries. If not all players are applying watermarks, it could be confusing for users, who might mistake untagged deepfakes for real content.

In addition to preventive measures, there are also detection technology solutions. Initial detection tools focused primarily on visual artifacts such as blended edges. More recently, deep learning techniques have capitalised on the large amounts of real and fake data available online. However, since training detection algorithms depends on fake data created by generation tools, deepfake detectors lag behind generators, according to experts. Conversely, the development of detection algorithms provides direct feedback to generation algorithms on what makes deepfakes detectable and can encourage generators to bypass detection. Moreover, AI providers are not required to make the complete AI's training data public, therefore there might not be independent oversight of its appropriateness or legality.

Some AI companies are taking measures to mitigate the impact of deepfakes. For instance, OpenAI has launched an initiative to detect deepfakes with a tool capable of distinguishing between AI-generated images and real ones. Other platforms use voice analysis tools to detect audio deepfakes or track the origin of content using blockchain technology. In this context, there are many events emerging around the world for developers to launch tools and technologies to detect them, such as the global Deepfake Detection Challenge.

So far, various detection approaches utilising machine learning, forensic analysis and hybrid techniques suffer from numerous limitations. A review analysis highlights the challenges of detecting synthetic voices in video-based deepfakes. Furthermore, actionable directions for future research and policymaking are proposed, such as fostering adaptive detection algorithms, promoting interdisciplinary collaborations and enhancing public awareness to manage the dual-edged nature of deepfake technologies. Insights into legislative efforts and ethical AI design principles are discussed to balance innovation and regulation.

Assessing the scale of the issue is difficult due to underreporting, a lack of reliable statistics and contaminated training datasets. In late 2023, an academic investigation found hundreds of images of child sexual abuse in an open dataset used to train popular AI text-to-image generation models such as the generative AI platform Stable Diffusion, which is widely used.

This may make determining the appropriate intervention difficult. Current detection methods and watermarking techniques, while progressing, show mixed results and face persistent technical challenges. No single robust solution currently exists to detect and reduce the spread of harmful AI-generated content. The rapid advancement of AI technology often outpaces detection methods, highlighting the potential limitations of relying solely on technical and reactive intervention.

One way to promote the protection of children when using generative AI tools is to provide them with age-appropriate AI and 'safe by design' tools. Age-appropriate AI ensures respect for children's cognitive development while 'safe by design' minimises misuse of generative AI to harm children. Some tech companies have committed to 'safe by design' principles and to training generative AI to proactively prevent generative models from generating CSAM. According to research, AI developers play a key role in combating CSAM generated using generative AI, and effective prevention measures may vary depending on whether the AI models are open-source or closed-source.

Legislation

The threat posed by misuse of deepfakes concerns countries globally, and legislation develops slower than technology, but it is emerging. China, for example, has mandated that all AI-generated content be watermarked for the purpose of combating AI disinformation, but not for other purposes. The UK has recently updated its Online Safety Bill to provide better protection for victims of deepfake abuse, making it illegal to create simulated child abuse, not just share it. For instance, a man was prosecuted under child sexual abuse law for transforming everyday photos of real children into CSAM. Some states in the US have also amended their penal codes to prohibit the creation, possession and distribution of deepfakes depicting minors in sexual contexts.

In the EU, although deepfakes are not currently banned under the legal framework, the AI Act recognises that synthetic content can pose a risk to the integrity of information and trust in it due to the difficulty of distinguishing between human-generated and synthetic content. Consequently, it established transparency obligations to label such content. Article 50(4) of the AI Act sets specific obligations for labelling deepfake content with the goal of ensuring transparency and preventing manipulation of the public. However, there are no definitive guidelines on how to label such content, and there are exceptions to the labelling obligation. For example, the obligation may not apply if AI-generated content has undergone human review or editorial control. The EU is in the process of developing codes of good practice on the labelling of AI-generated content by AI providers, expected to be published by August 2025.

The Digital Services Act (DSA) aims to provide a safer online environment by implementing measures to combat illegal content and harmful activities online. For instance, it set outs obligatory mitigation measures for large online platforms to address risks such as online harm to children.

Directive 2024/1385 on combating violence against women and domestic violence also addresses non-consensual images generated with AI, providing victims with protection[3] from deepfakes.

The proposal for a recast of the Children Sexual Abuse Directive aims to expand the definition of criminal offences to include CSAM in deepfakes. This presents an opportunity to tackle this growing threat to children, a move that has already been endorsed by the Civil Liberties Committee in the European Parliament. Some countries, like Spain, are already doing this at national level.

Civil law provides several options to challenge the use of deepfakes, including through trademarks and through lawsuits for copyright and privacy infringements, harassment or defamation.

Education

Education and awareness will have a key role in understanding the impacts of deepfakes and will contribute to making online users less prone to manipulation. Experts recommend checking certain details in generative AI output to distinguish a deepfake, such as a person's eyes, eyebrows and shadows, as well as their hair and movements in the video.

According to a report, children who participated in class discussions about generative AI were more likely to check the veracity of AI-generated output. However, the report also showed that 60 % of teenagers either attended schools without a policy on the use of generative AI or were unsure of their school's approach to generative AI. Moreover, half of the respondents in another report stated that they had not received training on deepfakes or, if they had received it, it was poor.

UNICEF emphasises that, even if the data about how children are using generative AI are limited, studies show that children are more likely to use it than adults. Therefore, it calls for addressing citizens' AI literacy. AI literacy will have an essential role in ensuring that children have the minimum AI skills to comprehend its threats. While AI literacy is more commonly taught at secondary schools and universities than at primary schools, the European Commission and the Organisation for Economic Co-operation and Development (OECD) are working on an AI literacy framework. The final version will be released in 2026 to support educators in integrating AI literacy into their teaching practices. Many experts suggest that, to prevent further harm, schools should update their codes of conduct, sexual harassment policies and cyberbullying policies to explicitly prohibit the creation and distribution of deepfake sexual imagery. Hence, experts highlight that student codes of conduct focused on cyberbullying can be used to tackle issues relating to deepfakes.

Next steps

Globally, UNICEF evaluated several national AI strategies and concluded that, while many countries are developing AI initiatives and guidelines, they must prioritise equitable strategies that will ensure the protection of children's rights and enable children to develop strong AI competencies at school.

The Lanzarote Committee published a declaration on protecting children against sexual exploitation and sexual abuse facilitated by emerging technologies, calling for protecting children against AI-generated materials. Stakeholders at the WeProtect organisation are calling for technology-neutral legislation that criminalises emerging threats such as AI-generated abuse, sextortion and deepfakes.

Some are also calling for holding legally accountable those who create and distribute harmful content, promoting digital literacy about the implications of harmful internet trends and implementing robust content moderation tools to better police images and prevent further victimisation through exposure.

This highlights the need for legal action and global cooperation to develop the tools and methods necessary to address deepfakes on a large scale and at a rapid pace. Additionally, there is a growing need to strengthen generative AI literacy for children, educators and parents, and to monitor indicators of children's online use, which are currently non-existent at the EU level. This should be accompanied by greater efforts by the industry and enhanced implementation of EU legislation such as the AI Act and the DSA.

MAIN REFERENCES

Niestadt, M., Protecting children online: Selected EU, national and regional laws and initiatives, EPRS, European Parliament, April 2025.

Marcelin, T., AI and copyright: The training of general-purpose AI, EPRS, European Parliament, April 2025.

Negreiro, M., Children and generative AI, EPRS, European Parliament, February 2025

Negreiro, M., Combating child sexual abuse online, EPRS, European Parliament, November 2024.

Murphy, C., Cyberbullying among young people: Laws and policies in selected Member States, EPRS, European Parliament, June 2024.

ENDNOTES

[1] Among them face swapping, reenactment, diffusion-based deepfakes, audio-driven facial animation, style-based generator for faces, and audio generation models.

[2] Category A in the UK, which includes CSAM that contains penetrative sexual activity, bestiality, or sadism.

[3] This directive, among others, requires Member States to criminalise the sharing of intimate material, meaning material depicting the intimate parts of a person or a person engaged in sexual activities, without the permission of the depicted person. This includes the creation of such intimate content through deep fakes and their dissemination. However, the prohibition on creating and sharing deepfakes only applies to deepfakes depicting a person engaged in sexual activities. According to this Directive, Member States must envisage measures for the prompt removal of such content or disabling access to such content.

DISCLAIMER AND COPYRIGHT

This document is prepared for, and addressed to, the Members and staff of the European Parliament as background material to assist them in their parliamentary work. The content of the document is the sole responsibility of its author(s) and any opinions expressed herein should not be taken to represent an official position of the Parliament.

Photo credits: © OleCNX / Adobe Stock.

eprs@ep.europa.eu (contact)

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

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

http://epthinktank.eu (blog)