Deloitte.

The rise of deepfakes: What digital platforms and technology organizations should know.

January 2025

Contents

Executive summary

Driven by recent advances in Generative AI tools, the proliferation of deepfake content on social media platforms has become a recent phenomenon, having grown 550% between 2019 and 2023[1]. This exponential growth has sparked heightened concerns from individuals, organizations, and governments worldwide, with the World Economic Forum calling deepfakes and disinformation as one of the key global risks in 2024[2]. The availability of Generative AI tools creates scaled opportunities for bad actors to create deepfake content and exploit vulnerabilities in digital platforms that lack preparedness for this new type of risk. Digital platforms should adopt a multi-faceted approach to assess and mitigate the risks to their users posed by deepfakes. This includes identifying potential harms to their users and supporting those users accordingly, assessing where existing processes and controls (such as those for login and account verification) may be impacted by deepfakes, and continuously evaluating detection tools and capabilities.

The purpose of this paper is to explore the associated risks due to deepfakes and how technology organizations, including digital platforms, can work to address these risks. It also delves into current regulatory considerations in the United Kingdom, European Union, and United States, and discusses practical prevention and detection mechanisms for digital platforms and organizations that use digital platforms for advertising and other content. Key points include the following.

- **Challenges and risks:** One of the most discussed concerns with deepfakes on digital platforms is the application to adult imagery and child sexual abuse material (CSAM), constituting a significant invasion of privacy and harassment. Other risks include disinformation, where deepfakes are being used to falsely depict public figures making statements or engaging in activities that never occurred, and fraud and scams, where deepfakes have been used to trick consumers and businesses to provide login information to bad actors, leading to account takeover, or to convince consumers and businesses to transfer funds to bad actors, including for fundraising or investment scams.
- **Regulatory obligations for risk assessments:** Digital platforms and technology organizations should conduct ongoing assessments of the types of risks perpetuated by deepfakes. The EU Digital Services Act specifically mandates those digital platforms classified as Very Large Online Platforms and Very Large Search Engines perform an annual systemic risk assessment to assess the key risks that their services may pose, including those of deepfakes and other inauthentic behaviors. Similar risk assessments will need to be performed by digital platforms with users in the United Kingdom under the UK Online Safety Act. The EU Artificial Intelligence Act also requires developers of general-purpose AI models to conduct a systemic risk assessment starting in August 2025. The European Parliament has set out one framework for inventorying these risks, categorizing the risks into psychological harms, financial harms, and societal harms.
- **Mitigation strategies:** Common mitigation include the detection and labeling of deepfake content, guardrails to prevent Generative AI tools from creating deepfake content (especially in high risk areas like political contexts), reporting and removal of deepfake content, and providing resources and support for users to identify deepfake content and for those who were affected by it accordingly.
- **Technology resources:** In addition to tools utilized for the detection of deepfake content, digital platforms are employing content labeling and provenance techniques, such as fingerprinting, digital signatures, and watermarks, to support the traceability and authenticity of the content accordingly. There are working groups to facilitate open source standards for these technologies.
- **Regulatory responses:** The regulatory response to deepfakes varies considerably across different regions, reflecting the complexities and challenges posed by this rapidly evolving technology. It additionally differs based on whether the organization is creating or disseminating content. In the United Kingdom, there is a new law under the Criminal Justice Bill criminalizing the creation of sexually explicit deepfakes, and Ofcom, the UK's Communications Regulator, indicated that it plans to assess merits and interventions for deepfakes in 2025. In the European Union, the EU Artificial Intelligence Act requires creators of Generative AI to label their content, preventing misinformation and protecting individuals, and the Digital Services Act mandates annual systemic risk assessments for Very Large Online Platforms and Very Large Search Engines. In the United States, there are a mix of state-level laws and federal initiatives, including California's legislation banning deepfakes in elections and criminalizing AI-generated child sexual abuse images, Texas criminalizing deceptive videos intended to influence elections, and the White House Executive Order on AI focuses on labeling, detecting, testing, and auditing synthetic content.

Introduction

A new type of risk – An illustrative example

A day ahead of a major earnings release, a video seemingly featuring an executive of a major corporation appears on multiple digital platforms. In this video, the executive discusses how the corporation significantly missed its earnings targets. The video is widely shared within only a few hours and generates discussion and analysis from users across those platforms, as well as news media. The stock markets react to the apparent news, and the corporation's stock price is impacted. Multiple digital platforms take down the video, recognizing it as a deepfake created by bad actors as a form of disinformation and stock price manipulation. But there is already impact to the reputation and finances of the corporation.

This raises questions – should the deepfake have been detected earlier? Did the digital platforms take appropriate precautions to address disinformation? How does the digital platform demonstrate its ability to do so?

Section 1 – Deepfakes

1.1. What are deepfakes?

Deepfakes are a type of synthetic media created using artificial intelligence (AI) that manipulates or generates text, documents, visual, or audio content. Although the majority of Generative AI content uses are legitimate, as further discussed below, deepfakes are differentiated due to their intent to deceive or demean. The term "deepfake" itself is derived from "deep learning," a subset of AI that uses algorithms to analyze and generate data, and "fake," reflecting the false nature of the content produced.

The concept of deepfakes first gained notoriety in 2017 when an online user shared realistic but fabricated inappropriate videos featuring well-known celebrities. Since then, the scope of deepfakes has expanded significantly, encompassing not just fake videos but also synthetic audio, images, and documents that are increasingly indistinguishable from real ones.

1.2. What are other uses of AI-generated content?

There are legitimate and innovative applications in sectors such as entertainment, education, marketing, and security. This includes the following examples, which will likely expand as the use of Generative AI grows:

- **Entertainment and media:** Generative AI is being used by the entertainment industry to enable digital recreations of actors, personalized marketing content, and more immersive video game experiences.
- **Education:** Generative AI is being used to create interactive language learning tools, bring historical figures to life, and enhance training simulations with realistic scenarios.
- **Corporate and marketing:** Businesses are using Generative AI to develop engaging training materials, create customized marketing campaigns, and improve customer service interactions through lifelike virtual assistants.
- **Security and fraud prevention:** Generative AI can be used for testing and improving biometric systems, as well as training law enforcement with realistic simulations.
- **Health care:** In medicine, Generative AI is applied to simulate patient interactions for training purposes and to create virtual avatars that support therapeutic treatments.
- **Art and creativity:** Artists and musicians are exploring Generative AI to push the boundaries of creativity, producing innovative digital art and recreating iconic voices in new musical works.

1.3. How are bad actors using deepfakes on digital platforms?

Human exploitation and harassment

One of the most discussed concerns with deepfakes is its application to adult imagery and child sexual abuse material (CSAM). In a study of over 95,820 deepfake videos online on digital platforms in 2023, one security research agency found that 98% of deepfakes were adult in nature, with 99% of these deepfake videos targeting women[1]. This application of deepfakes constitutes a significant invasion of privacy and a form of harassment.

Disinformation, public safety, and public health

Deepfakes are increasingly being used in a political context as a form of disinformation. Political deepfakes can involve the creation of videos or audio recordings that falsely depict public figures making statements or engaging in activities that never occurred. These fabricated media can be used to spread misinformation, manipulate public opinion, and potentially influence election outcomes. Slovakia serves as a cautionary tale, where a deepfake audio surfaced just two days before the 2023 parliamentary elections and seemingly featured one of the leading candidates discussing electoral fraud with a prominent journalist.[2] According to the World Economic Forum's Global Risk Report 2024[3], misinformation and disinformation driven by deepfakes are ranked among the most prominent global risks to democracy in the next two years, as more than 60 countries worldwide held elections in 2024.

Deepfakes and disinformation have also surfaced during public health emergencies such as the COVID-19 pandemic. Researchers found that disinformation on digital platforms elicited stronger reactions and led to higher circulation as compared to legitimate information about public health concerns[4].

Financial markets

Given the velocity of financial markets, disinformation can have an immediate and substantial impact on any one corporation, or the financial markets more broadly. The immediate effect of disinformation has already been observed: This has resulted in impacts to financial markets, which demonstrates how even information that is easily verifiable as false can cause disruption – as deepfakes featuring the images of individual executives or corporations may surface, the response and recovery may not be as immediate.

Fraud and scams

Deepfakes are increasingly used to perpetrate financial fraud. Over the last two years, there have been many examples of large financial services firms that have transferred funds, thinking that senior executives or customers have expedited approval of the transfer. The voice on the call and images on the videos were generated using deepfake technology, mimicking the executives' and customers' voices to persuade the treasury departments to authorize the transfer. Over half of C-suite and other executives (51.6%) expect an increase in the number and size of deepfake attacks targeting their organizations' financial and accounting data—otherwise known as deepfake financial fraud—during the next 12 months, according to a recent Deloitte poll.[5]

Similar technologies are being used to trick consumers and businesses to provide login information to bad actors, leading to account takeover, or to convince consumers and businesses to transfer funds to bad actors, including for fundraising or investment scams. Deloitte's Center for Financial Services estimates that deepfakes could enable fraud losses in the United States to reach US$40 billion by 2027, up from US$12.3 billion in 2023, a compound annual growth rate of 32%[6].

Account authenticity

As more digital platforms adopt identity verification controls, such as those to verify the age or identity of a user, Generative AI tools used to create deepfakes provide bad actors with the ability to easily manipulate images or create synthetic identities. This creates a potential vulnerabilities in an organization's ability to identify trusted individuals or celebrities, as well as protect teens and children with specialized accounts or features.

To the extent that digital platforms have payments products and support Know-Your-Customer due diligence, deepfake documents enable bad actors to easily simulate synthetic forms of identification, such as driver's licenses and passports.

1.4. How are advances in Generative AI leading to increases in deepfakes?

Deepfakes rely on advanced machine learning techniques, including neural networks, generative adversarial networks (GANs), variational autoencoders (VAEs), and text-to-speech (TTS) technologies. Advances in Generative AI technology have increased the availability and accessibility of these techniques, including for bad actors.

Photo and video editing software, which has been available for legitimate purposes for over two decades, are now adopting Generative AI capabilities. While many organizations hosting these tools are developing guardrails and conducting testing related to safety use cases, bad actors continue to evolve in their sophistication to evade these guardrails. Researchers in a 2024 study[7] identified that bad actors leverage both "uncensored" models (e.g., those models that can readily generate harmful content given lack of content filtering) and "jailbreak" prompts (e.g., adversarial inputs) on otherwise legitimate available models. These tools and instructions are readily accessible for purchase and use via marketplaces and forums. Moreover, the researchers identified that prices associated with these tools and instructions generally were significantly lower than those of traditional malware tools[8].

A variety of malicious Generative AI services have emerged as well. These include DarkGPT and EscapeGPT, each of which produced content that demonstrated high evasion of detection by anti-virus and other detection tools in the aforementioned study. As another example, WolfGPT is capable of producing phishing emails which allow fraudsters to evade phishing email detectors.

Section 2 – How can digital platforms mitigate these risks?

2.1. What types of risks are digital platforms identifying?

Digital platforms and technology organizations should conduct ongoing assessments of the types of risks perpetuated by deepfakes. The EU Digital Services Act specifically mandates digital platforms classified as Very Large Online Platforms and Very Large Search Engines to perform an annual systemic risk assessment to assess the key risks that their services may pose, including deepfakes and other inauthentic behaviors[9]. Similar risk assessments will need to be performed by digital platforms with users in the United Kingdom under the UK Online Safety Act[10]. The EU Artificial Intelligence Act also requires developers of general-purpose AI models with systemic risk to carry out a systemic risk assessment of the risks of their models at the EU level starting in August 2025[11].

The European Parliament has set out one framework for inventorying these risks[12], categorizing the risks into psychological harms, financial harms, and societal harms, as listed below.

Psychological harm

- Sextortion
- Defamation
- Intimidation
- Bullying
- Undermining Trust
- Child sexual abuse deepfake material
- Gender-based violence
- Harassment
- Stalking (using various profiles with deepfake personas to stalk an individual)
- Invasion of privacy
- Online sexual harassment
- Non-consensual pornography

Societal Harm

- News media manipulation
- Damage to economic stability
- Damage to justice system
- Damage to scientific system
- Erosion of trust
- Damage to democracy
- Manipulation of elections
- Damage to international relations
- Damage to national security

Financial Harm

- Extortion
- Identity theft
- Fraud
- Stock-price manipulation
- Brand damage
- Reputational damage

As part of their risk assessment, organizations should gather information—including data and metrics—related to the prevalence of deepfake content on the platform. Examples of this may include the volume of potential deepfake content; volume of potential violations of deepfake content through their existing processes (e.g., identity verification or call center); and volume of potential deepfake content identified through their existing reporting mechanisms. Deepfake detection tools can assess image and video content for authenticity, but analysis will need to be performed to determine whether the content constitutes deception or demeaning purposes.

After identifying these risks, digital platforms should evaluate the coverage, design, and operating effectiveness of their active mitigation measures, as further listed in the section below.

2.2. What types of mitigations can digital platforms utilize?

Given the variety of risks, there are a wide variety of risk mitigations and controls that digital platforms can undertake. As an example, the European Commission presents a framework for local governments and their regulatory agencies to respond to the risks.[13] Digital platforms can adopt a similar framework for their risk mitigations, as illustrated below.

Technology dimension

A. This would include the ability for the digital platform to use a variety of signals and information—including deepfake detection mechanisms, watermarks, metadata, fingerprints, and other cryptographic methods—to detect and label potential deepfake content accordingly.

B. Where Generative AI models leverage data sources for information and training, confirm that these are reliable sources and provide sources of information for users to perform additional research, as needed.

Creation dimension

A. To the extent that a digital platform or similar service allows users to create content using Generative AI tools, this would include the ability to ban certain types of content altogether, such as content involving political candidates or those involving nudity.

B. This additionally would include guardrails to prevent AI-generated content from inadvertently creating content in the areas above, as an artifact of model risk and even if it was not prompted to do so.

C. This also includes the detection and tracking of bad actors who are repeatedly trying to create this content, banning these users from the platform, and escalating these bad actors to law enforcement accordingly.

Circulation dimension

A. Acknowledging the potential spread of disinformation content, this would include the digital platform's ability to respond once the content has been identified as potential deepfakes through the form of taking down content, down-ranking content so that it does not appear in search or recommended feeds, sharing external signals on potentially deepfake content, and supporting the appeals process. This can include the evaluation of recommender systems to provide a reduction of disinformation prominence.

B. Facilitate collection of user reporting of potential deepfake content.

C. Consolidate a diverse set of signals—including feedback from deepfake detection tools, device and network information from the user who uploaded the content, and the user's history of nefarious content, if applicable—in order to analyze and proactively prevent content from being uploaded to the digital platform.

D. Empower fact-checkers with the tools to quickly evaluate potential content for detection of their veracity and their potential harm.

Target dimension

A. Recognizing the negative psychological effects on victims of deepfakes, this would include providing cyberbullying support resources and strengthening data protection around key data sources used for the creation of deepfakes.

Audience dimension

A. This would include resources to the audience to allow users to identify potential deepfake content, including labeling trustworthy sources, providing additional context related to high risk topics (e.g., elections or public health) and supporting media literacy programs.

B. Digital platforms will need to account for the unique cultural and political contexts of each country. Since disinformation vary across regions due to individual political and cultural factors, platforms should adopt a localized approach and evaluate the efficacy of their mitigations (including deepfake detection) using local languages and contexts accordingly.

2.3. How can digital platforms detect deepfakes?

An important capability in a deepfake mitigation strategy is the ability to identify potential deepfake content, which can be performed through either manual or automatic means.

Manual detection relies on a human to inspect a video or an audio recording to assess the veracity of the recording. Humans can notice an uncanny appearance of a character's movements or sounds that give out clues that it is AI-generated. However, ongoing advancements to deepfake technologies, combined with the need to scale prevention and detection mechanisms, can limit the usefulness of manual detection for digital platforms.

Automatic deepfake detectors usually include a combination of systems that are AI-based themselves. These include the following capabilities:

- Advanced models (e.g., neural networks) trained to differentiate between authentic and synthetic media
- Temporal inconsistencies and visual artefacts
- Video and voice liveness detection
- Facial recognition
- Facial feature analysis
- Metadata analysis

Many digital platforms and media organizations utilize a combination of manual and automated detection measures as part of their misinformation and disinformation controls. For example, some news agencies are implementing fact-checking and video verification through a combination of automated technologies and supplementary research.

However, given the growing sophistication of deepfake tools, there are numerous challenges with deepfake detector systems. As deepfake detection systems improve, bad actors are expected to continuously improve detection evasion techniques. Bad actors are also adept at layering in many different evasion methods to avoid deepfake detection.

Digital platforms additionally should recognize that there are many legitimate uses of Generative AI and content generation, and it can sometimes be hard without additional context to understand the intent of synthetic content. Digital platforms should utilize available signals, including the behavior of the actors uploading the content, to identify if it constitutes takedown and further enforcement.

Digital platforms and other technology organizations should recognize that a detection strategy is not a singular tool or control but a multi-faceted approach that requires proactive monitoring, analysis, agility, and evolution. Given the fast-moving nature of the AI and Generative AI space, there is not a singular benchmark for a robust framework; as a result, deepfake attacks will likely continue to rapidly evolve.

2.4. How can digital platforms leverage content labeling and provenance?

As deepfakes gain prominence worldwide, the adoption of watermarking and content provenance as tools to detect and trace their origin is growing. Various governments are encouraging these methodologies to combat the proliferation of deepfake content. For example, under the transparency requirements of the EU Artificial Intelligence Act, there will be new labeling requirements for organizations generating or disseminating synthetic content.

Content provenance refers to the documented history of the origin of digital content. Content provenance should be able to provide details about the content's creation, modification, ownership, and dissemination across its life. Content provenance increasingly serves an important role in promoting trust, authenticity, and transparency in the digital ecosystem, helping to safeguard the integrity of information and protect against deception and manipulation.

Four examples of technologies for establishing media provenance are metadata, fingerprinting, digital signatures, and watermarking. Individually, these technologies have vulnerabilities; but when combined, they can mutually reinforce each other to create a more robust system for media provenance:

Metadata

Metadata carries information about the media file, such as its origin, date of creation, and any modifications it has undergone. However, metadata can be stripped or altered, making it vulnerable to tampering.

Fingerprinting

Fingerprinting (also known as hashing) is a method of recording a base copy of media (such as pixels on an image or video), to be used later to assess if there are fake or altered versions in circulation.

Digital signatures

This method involves creating a unique digital signature for media content using hashing, and using this digital signature to track its authenticity over time. This can be an effective method to identify deepfakes even after original content has been transformed, but it relies on maintaining a trusted repository of these signatures.

Watermarking

Watermarking embeds signals in the media that can later be recovered to verify its authenticity. While watermarks can persist through some transformations, they can also be stripped or spoofed, limiting their effectiveness.

Journalists, news editors, and non-governmental organizations are developing content provenance standards. One of the leading organizations within the content provenance area is the Content Authenticity Initiative (CAI). The CAI was founded in November 2019 by a variety of technology and news media organizations, to provide an open-source software development kit that is compliant with the Coalition for Content Provenance and Authority (C2PA) specification for content provenance. Several digital platforms have additionally signed on to participate in the C2PA.

2.5. What are additional considerations when licensing data to third parties?

To the extent that digital platforms license their data, including for training of external AI systems, they may be asked by their customers to demonstrate the reliability of the content provided. This extends to both content provided by users as well as advertisers, and especially includes where AI-generated content contains copyrighted media. As part of content provenance commitments, including those as part of C2PA, digital platforms may need to demonstrate how copyrighted media are being used, how it is not being manipulated or transformed during transmittal, and which third parties are accessing it. There also is a risk of the licensed third party improperly misusing a digital platform's data, including for the creation of deepfake content. Digital platforms should evaluate third-party relationships and include proactive controls – including contractual agreements on data use, establishing terms of services, and monitoring for misuse accordingly.

To the extent that this content has been developed through Generative AI tools, and especially when this deepfake content uses brand names, logos, or trademarks, it also impacts the reliability of the content being monetized.

Section 3 – Regulatory regime for deepfakes

The regulatory response to deepfakes varies considerably across different regions, reflecting the complexities and challenges posed by this rapidly evolving technology. It additionally differs based on whether the organization is creating or disseminating content.

United Kingdom

The UK's Communications Regulator (Ofcom) recently introduced a Discussion paper on tackling deepfakes[14]. They identified three types of deepfakes: 1. Deepfakes that demean, 2. Deepfakes that defraud and 3. Deepfakes that disinform. The paper outlined four key areas of intervention to be applied by different actors at different stages in the technology supply chain: 1. Prevention, 2. Embedding, 3. Detection and 4. Enforcement. During 2025, Ofcom plans to assess the merits and limitations of deepfakes areas of intervention, including the hashing and forensics techniques for deepfakes. Subsequently, Ofcom will consider whether the measures would be included in the Codes of Practice or Guidance linked to the UK Online Safety Act[15].

The UK government also introduced a new law under the Criminal Justice Bill that criminalizes the creation of sexually explicit deepfakes, even if there is no intent to share them[16]. This law builds on previous measures under the Online Safety Act, which made it illegal to share non-consensual intimate images. The penalties under this new offence include unlimited fines or criminal records.

European Union

The European Union has taken proactive steps to regulate deepfakes, though challenges remain. The EU Artificial Intelligence Act, for instance, requires creators of Generative AI to label their content accordingly, making it clear that the media has been manipulated. This regulation is intended to prevent the spread of misinformation and protect individuals from the harm that deepfakes can cause.

Anticipating the potential for interference, the European Union published 2024 Digital Services Act (DSA) Election Guidelines for Very Large Online Platforms and Very Large Online Search Engines that outlined the potential mitigation techniques to address the risks of deepfakes and disinformation. Deepfakes will need to be assessed during the digital platforms' risk assessments and will need to be considered across sections of the DSA including protection of minors, as well as removal and downgrading of content. Because deepfakes can be used to impact both disinformation and hate speech, digital platforms will also have to consider deepfakes within the context of the EU Codes of Conduct and Codes of Practice, such as the Strengthened Code of Practice against Disinformation or the Code of Conduct Countering Illegal Hate Speech.

These guidelines need to be considered in conjunction with other local or international regulations and agreements, such as the Tech Accord to Combat Deceptive Use of AI in 2024 Elections signed during the Munich Security conference[17]. There are additional country-specific laws, including Germany's Network Enforcement Act or France's Law on the Fight Against Manipulation of Information, where deepfake risks would apply. This means that deepfakes in the EU are being regulated through a combination of new and existing regulation, which organizations will have to navigate in parallel.

It is important to note that deepfake detection systems used by law enforcement agencies fall into the category of 'high-risk AI' per the EU AI Act, as they could pose a risk to the individual rights and freedoms of the individuals. Therefore, the deepfakes detection systems will have to undergo risk assessments and comply with strict data governance and data management processes.

United States

In the United States, the response to deepfakes includes a mix of state-level laws and federal initiatives. This includes the following state-level laws, among others:

- California has passed legislation banning the use of deepfakes in elections, requiring social media platforms to remove deceptive material close to election dates, and criminalizing AI-generated child sexual abuse images.[18]
- Texas, among other states, have criminalized deceptive videos with an intent to injure a candidate or influence the outcome of an election.[19]
- Florida criminalized deepfake images to portray an identifiable minor engaged in sexual conduct[20]. Similarly, Louisiana and South Dakota criminalized deepfakes involving minors engaging in sexual conduct[21, 22].

At the federal level, the White House Executive Order on AI focuses on the development of tools and methods to label, detect, test and audit synthetic content[23]. The order also aims to prevent Generative AI from producing child sexual abuse material, among other harmful content.

The United States Senate has introduced the Protect Elections from Deceptive AI Act, which aims to prohibit the distribution of materially deceptive AI-generated audio or visual media relating to candidates for federal office, and for other purposes[24].

The United States Senate additionally has passed the Defiance Act 2024, which aims to improve rights of those affected by non-consensual intimate imagery[25].

The United States House of Representatives has introduced the DEEPFAKES Accountability Act, which aims to protect individuals as well as national security against the threats of deepfakes[26]. The United States House of Representatives additionally has introduced the Protecting Consumers from Deceptive AI Act, which sets up task forces to facilitate and inform the development of technical standards and guidelines relating to the identification of content created by Generative AI[27].

Section 4 – How can digital platforms get started?

Even with the evolving nature of deepfakes, there are some initial actions that digital platforms can introduce to mitigate the risks of harms linked to deepfakes. Below we have outlined actions that digital platforms may consider implementing:

Update platform policies 1

Clear and enforceable guidelines on deepfakes should be established, outlining what constitutes a violation and the consequences for users and advertisers who distribute such content.

Adopt advanced detection technologies 2

Digital platforms should deploy AI-powered tools to detect and flag potential deepfakes. Signal from a deepfake detection tool can be combined with data from the users' device, network, and historical behaviors.

Assess the prevalence of deepfake content 3

Using deepfake detection tools and feedback received from their users, digital platforms can assess the prevalence of different types of deepfakes to inform future investments in capabilities and controls.

Evaluate content moderation 4

Due to the velocity and apparent authenticity of deepfake content, the rise of deepfakes will pressure test a digital platform's content moderation practices. Digital platforms should identify and address vulnerabilities within their existing capabilities and controls, particularly in preparation of critical events where misinformation can spread rapidly.

Educate users 5

Digital platforms can launch campaigns to raise awareness about deepfakes, providing users with tools and tips to recognize manipulated media.

Simulate an event 6

One common way of evaluating the efficacy of mitigations is through red-teaming exercises with the relevant industry specialists, often with focus on a given use case, like the electoral process or child safety.

Adopt ongoing monitoring and vigilance 7

Generative AI capabilities are rapidly evolving, and bad actors will continue to find ways to use these tools for nefarious purposes. Adopt continuous evaluation of risks and schemes, as well as ongoing monitoring and training of the detection tools and capabilities that your organization is leveraging.

Collaborate with regulators and peers 8

Digital platforms should work together and with regulatory agencies to create and follow industry-wide standards for managing content authenticity and identifying deepfakes.

Support victims 9

Companies should offer resources for those affected by deepfakes, including reporting tools and support services to help mitigate the impact of deepfake harassment or abuse.

Section 5 – A call to action

The rise of Generative AI tools has significantly increased the potential for deepfake content. As these tools become more accessible, digital platforms should adopt strategies to detect and mitigate these threats. This includes vigilant monitoring, tailored risk management activities, and collaboration with peer institutions, law enforcement, as well as international and local agencies.

By taking these steps, digital platforms can demonstrate to their users and their regulatory agencies that they are responding effectively and building trusted communities.

How can Deloitte* help your organization?
As a leading provider of services to digital platforms and other technology organizations, we have a multidisciplinary team of risk and technology specialists to assist organizations in their response to new and evolving risks, including those amplified through deepfake technologies:

- **Assessing risks and aligning controls** – Identifying, quantifying, and assessing the risks posed to your organization through deepfakes is often one of the first steps in building a deepfake strategy. We can assist you in assessing the types of schemes enabled by deepfakes, identify the impact of those schemes, and communicate the outcome of these assessments to key stakeholders, including senior leadership and your boards of directors. This additionally can be used to help evidence compliance with regulatory obligations.
- **Technology enablement** – With the rise of deepfakes, there are a variety of ways organizations are leveraging technology to detect and respond to potential deepfake content in a proactive, scalable manner. We assist organizations with a variety of services related to technologies to help prevent and detect deepfakes, including the following areas:
- The assessment of evolving threat scenarios
- The evaluation of potential technology capabilities and vendors, including testing or "red teaming" of Generative AI tools for potential misuse
- Alignment of tools to threat vectors (e.g., customer call centers, employee video conferencing tools)
- Configuration of tools and the testing of their capabilities, including those for deepfake content detection, takedown, and enforcement against bad actors
- Simulations to test how your organization and its systems would respond to a deepfake incident and understand ways to mitigate risk incurred by the use of deepfakes.
- **Incident response** – To the extent that deepfakes were involved in an event on your digital platform, at your enterprise, or when a customer was using one of your services, we can assist with your organization's response to the incident:.
  - Obtain information about potentially impacted areas, including those for financial or reputational risks:
    - Identify user(s) who uploaded the content, whether the content was scanned for potential deepfake signals, and—if it was scanned—assess why it was not detected and enforced upon
    - Assess whether the content is still in circulation, or if there are tangential forms of the content (e.g., video snippets or ongoing discussions)
    - Conduct an impact assessment of the volume and type of users who were exposed to the content
    - Identify whether user, customer, or employee information was transmitted during the deepfake incident. Social engineering and phishing through deepfakes can lead to compromised credentials and unintended access to enterprise systems to facilitate fraudulent or malicious activity. This can include, but may not limited to, your customers' payment information, access to sensitive information (e.g., personally identifiable information or protected health information), your organization's procurement and treasury operations, and theft of your organization's intellectual property.
  - Fact-finding and support during litigation

Contact Us

Lead Author

Lenka Rueda Molins
Associate Director
Deloitte North and South Europe
lmolins@deloitte.co.uk

UK contacts

Nick Seeber
Partner
Deloitte North and South Europe
nseeber@deloitte.co.uk

Joey Conway
Partner
Deloitte North and South Europe
jconway@deloitte.co.uk

Scott Bailey
Director
Deloitte North and South Europe
scottbailey@deloitte.co.uk

US contacts

Brendan Maggiore
Senior Manager
Deloitte Transactions and Business Analytics LLP
bmaggiore@deloitte.com

Matt Galek
Specialist Master
Deloitte Consulting LLP
mgalek@deloitte.com

Mike Weil
Managing Director
Deloitte Financial Advisory Services LLP
miweil@deloitte.com

Endnotes

1 Security Hero. 2024. 2023 State of Deepfakes. https://www.securityhero.io/state-of-deepfakes/.

2 Harvard Kennedy School Misinformation Review. 2024. Beyond the deepfake hype: AI, democracy, and "the Slovak case". https://misinforeview.hks.harvard.edu/article/beyond-the-deepfake-hype-ai-democracy-and-the-slovak-case/.

3 World Economic Forum. 2024. 4 ways to future-proof against deepfakes in 2024 and beyond. https://www.weforum.org/stories/2024/02/4-ways-to-future-proof-against-deepfakes-in-2024-and-beyond/.

4 Bahareh Farhoudinia, Selcen Ozturkcan & Nihat Kasap. 2024. "Emotions unveiled: detecting COVID-19 fake news on social media." Nature Humanities and Social Sciences Communications https://www.nature.com/articles/s41599-024-03083-5.

5 Half of Executives Expect More Deepfake Attacks on Financial and Accounting Data in Year Ahead – Press release | Deloitte US

6 Deepfake banking and AI fraud risk | Deloitte Insights

7 Advik Raj Basani and Xiao Zhang. 2024. GASP: Efficient Black-Box Generation of Adversarial Suffixes for Jailbreaking LLMs. https://arxiv.org/html/2411.14133v1

8 Zilong Lin, Jian Cui, Xiaojing Liao, and XiaoFeng Wang. 2024. Malla: Demystifying Real-world Large Language Model Integrated Malicious Services. https://arxiv.org/html/2401.03315v1

9 European Commission. The EU's Digital Services Act. https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en

10 Legislation.gov.uk. 2023. Online Safety Act 2023.https://www.legislation.gov.uk/ukpga/2023/50

11 EU Artificial Intelligence Act. 2024. Up-to-date developments and analyses of the EU AI Act. https://artificialintelligenceact.eu/

12 European Parliament Research Service. 2021. Tackling deepfakes in European Policy. https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf

13 European Parliament Research Service. 2021. Tackling deepfakes in European Policy. https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf

14 Office of Communications (UK). 2024. A deep dive into deepfakes that demean, defraud and disinform - Ofcom. https://www.ofcom.org.uk/online-safety/illegal-and-harmful-content/deepfakes-demean-defraud-disinform/

15 Office of Communications (UK). 2024. Deepfake Defences: Mitigating the Harms of Deceptive Deepfakes. https://www.ofcom.org.uk/siteassets/resources/documents/consultations/discussion-papers/deepfake-defences/deepfake-defences.pdf

16 GOV.UK. 2024. Government cracks down on 'deepfakes' creation - GOV.UK. https://www.gov.uk/government/news/government-cracks-down-on-deepfakes-creation

17 Munich Security Conference. 2024. A Tech Accord to Combat Deceptive Use of AI in 2024 Elections. https://securityconference.org/en/aielectionsaccord/

18 Governor of California. 2024. Governor Newsom signs bills to combat deepfake election content. https://securityconference.org/en/aielectionsaccord/

19 Zach Williams. Bloomberg Law. 2024. More States to Push Laws Banning AI Election Deepfakes in 2024. https://news.bloomberglaw.com/artificial-intelligence/more-states-to-push-laws-banning-ai-election-deepfakes-in-2024

20 Senate Bill 1798 (2022) - The Florida Senate. https://www.flsenate.gov/Session/Bill/2022/1798

21 Louisiana Senate. https://www.legis.la.gov/Legis/ViewDocument.aspx?d=1309797

22 South Dakota Attorney General. https://atg.sd.gov/OurOffice/Media/pressreleasesdetail.aspx?id=2505#gsc.tab=0

23 Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence | The White House

24 Text - S.2770 - 118th Congress (2023-2024): Protect Elections from Deceptive AI Act | Congress.gov | Library of Congress

25 Text - S.3696 - 118th Congress (2023-2024): DEFIANCE Act of 2024 | Congress.gov | Library of Congress

26 Text - H.R.3230 - 116th Congress (2019-2020): DEEP FAKES Accountability Act | Congress.gov | Library of Congress

27 Text - H.R.7766 - 118th Congress (2023-2024): Protecting Consumers from Deceptive AI Act | Congress.gov | Library of Congress

Deloitte.

Designed by CoRe Creative Services. RITM1974396