techUK

Discussing Deepfakes:

The opportunities and challenges of synthetic media technology

About techUK

techUK is the trade association which brings together people, companies and organisations to realise the positive outcomes of what digital technology can achieve. With around 1,000 members (the majority of which are SMEs) across the UK, techUK creates a network for innovation and collaboration across business, government and stakeholders to provide a better future for people, society, the economy and the planet. By providing expertise and insight, we support our members, partners and stakeholders as they prepare the UK for what comes next in a constantly changing world.

Contents

Executive Summary

This report explores the growing threat of deepfakes and the positive opportunities presented by synthetic media.

The report outlines some recommendations for a collaborative response from government, industry, and society. Effective measures require a multi-pronged approach encompassing legal frameworks, technological advancements, and public education.

The current legal framework attempts to tackle the negative impacts of deepfakes through a range of legislation, from recent measures in the Online Safety Act and amendments in the Criminal Justice Bill among others.

But real-time actions are what is crucial to both mitigate harm, both from deepfakes and to continue innovation in synthetic media.

Synthetic media is rapidly transforming how we interact with information, unleashing a new wave of creativity and innovation. It offers immense potential to revolutionise education, enhance accessibility, and empower innovative storytelling, from text to image generators and synthetic voice generation to AI avatars and personal assistants.

Deepfakes present challenges, categorised primarily into two risks: being deepfaked and being deceived by deepfakes. Their prevalence is only increasing, with there being a 550% rise in total deepfake videos online from 2019 to 2023. Deepfakes can be maliciously used to spread disinformation, commit fraud, infringe on privacy, weaken elections and national security and undermine trust and democracy. As deepfakes become more convincing, the urgency to mitigate their negative impacts grows.

As deepfakes become harder to detect with the naked eye, a combination of cross sector tools is necessary to combat them. Solutions such as detection and mitigation technologies play a key role. Cross-sector initiatives, such as the AI Elections Accord, also focus on collaborative efforts to prevent deceptive AI content. Watermarking and labelling techniques ensure media authenticity by embedding cryptographic data into images. Fact-checking tools play a crucial role in disproving deepfakes, while AI-driven identity verification enhances security in digital interactions. Finally, increasing media literacy through education is also essential for the public to discern authentic from manipulated content.

As we move forward, it becomes crucial for governments, industries, and society to collaboratively shape a comprehensive legal and policy framework and scale out solutions.

Any future steps should focus on three key areas:

1. Innovation

Online Safety Sandbox: An online safety sandbox which focuses on disinformation, misinformation and deepfakes could be used to create a vast library of synthetic media, including both real and manipulated content and train AI models to better identify deepfakes in the real world.

Encouraging innovation for synthetic media: Synthetic media should be included as a focus area within existing AI technology challenge funds to enable research and development (R&D) in applications across various sectors.

2. Standards and best practice

Accelerate Adoption of Content Provenance: C2PA should be supported across industry as current best practice in content provenance. This would aid the identification of authentic media and address rising threats of disinformation.

Support Deepfake Detection, Labelling and Fact Checking Tools: Ofcom should promote the existing best practice that companies could use to tackle the malicious uses of deepfakes.

3. Media Literacy and Collaboration

Promote National Media Literacy: The Government, relevant regulators and industry should collaborate to facilitate training sessions for journalists and professions likely to be targeted by deepfakes. Deepfake and disinformation focused media literacy should also form an integral part of children's education in schools.

International Cohesion: Given the global and multifaceted nature of the deepfake problem, addressing it requires a collaborative effort. Other states have begun to invest heavily in labelling and media literacy surrounding deepfakes, and it will be crucial to learn from any actions of best practice.

Introduction

The rise of synthetic media presents new opportunities such as expanding our creative abilities in the film, art and advertising fields and bringing personal Artificial Intelligence (AI) assistants. However, the ability to create and general media more easily has also ushered in new challenges, such as the rise in new approaches to creating deepfakes and in combating child sexual abuse material (CSAM), disinformation and fraud.

As pioneers in technology, techUK members are leading the charge both in the development of synthetic media tools that can support businesses and public services, as well as in tackling the malicious use of AI-generated content.

Synthetic media and deepfakes: What are they?

Synthetic media is an all-encompassing term to describe any type of content whether video, image, text or audio that has been partially or fully generated using AI or machine learning. Types of synthetic media can range from AI-written music, text generation such as OpenAI's ChatGPT, computer generated imagery (CGI), virtual reality (VR), augmented reality (AR) and voice synthesis.

Deepfakes are a specific subset of synthetic media that focus on manipulating or altering visual or auditory information, to create convincing fake content, ranging from images, audio and video. A common example of deepfake use is videos that replace one person's face with another.

The key distinction between synthetic media and deepfakes is that the latter typically involves creating content that appears to be real but is fabricated with the intent to deceive, whereas synthetic media involves generating content for creative or practical purposes without the purpose of deception.

The term "deepfakes" is derived from the fact that the technology involved in creating this style of manipulated content uses deep learning techniques. Deep learning is a subset of machine learning, where a model uses training data to develop skills for a specific task. The more robust the training data, the better the model gets.

Deepfakes can often be created with malicious intent to deceive viewers and are becoming highly realistic and difficult to distinguish from genuine recordings or images.

In practice, deepfakes take the following most common forms:

- Face re-enactment, where advanced software is used to manipulate the features of a real person's face;
- Face generation, where advanced software is used to create entirely new images of faces using images of data from many real faces, but which do not reflect a real person; and
- Speech synthesis, where advanced software is used to create a model of someone's voice

When deepfakes were first developed several years ago, their creation required a high level of skill in AI, training, and technology, along with advanced equipment and time. However, with the development of AI, deepfakes are quickly becoming almost as accessible as cheapfakes to the general population.

In late 2017, Motherboard reported on a video that had appeared online in which the face of actress Gal Gadot had been imposed on an existing pornographic video to make it look like Gadot was engaged in the acts depicted. Despite being a fake, the video quality was good enough that a casual viewer could be fooled. An anonymous user of the social media platform Reddit, who referred to himself as "deepfakes," claimed to be the creator of this video. This was one of the first examples of an AI generated deepfake that gained significant attention.

How do deepfakes work?

Generative Adversarial Networks (GANs)

Some of the most common deepfake techniques leverage Generative Adversarial Networks (GANs). GANs involve two neural networks that are pitted against each other in an adversarial process. Data that represents the type of content to be created is fed to the first network so that it can 'learn' the characteristics of that type of data. The generator then attempts to create new examples of that data which exhibit the same characteristics of the original data. The second network is the discriminator, or adversary, which evaluates whether they are real or fake and 'learns' to identify the characteristics of that type of data. The adversary network attempts to detect flaws in the presented examples and rejects those which it determines do not exhibit the same sort of characteristics as the original data – identifying them as "fakes." These fakes are then 'returned' to the first network, so it can learn to improve its process of creating new data. With each iteration, the generator refines the output. This back and forth continues until the generator produces fake content that the adversary identifies as real.

The first practical application of GANs was established by Ian Goodfellow and his coworkers in 2014[1], when they demonstrated the ability to create realistic synthetic images of human faces. While faces are a popular subject of GANs, they can be applied to any content. The more detailed the content used to train the networks in a GAN, the more realistic the output will be.

Diffusion Models

Conversely, diffusion models are a type of deep generative model that add noise to the training data and then reconstruct the data by reversing this process. While GANs are like a legal debate where each side sharpens the other's skills, diffusion models take a broad idea and refine it into something more detailed. Whereas GANs excel in creating faces and expressions, diffusion models can create realistic textures and patterns, making them a useful tool to fabricate convincing environments or contexts. Diffusion models are particularly useful in the image domain and are behind tools such as Dall-E or Stable Diffusion.

Cheapfakes: The Deepfake Predecessor

Deepfakes are not new. Society has been able to manipulate media for decades. In the lead-up to UK elections in 1983, excerpts from speeches by Margaret Thatcher and Ronald Reagan were spliced together to create a fake telephone conversation, in which both leaders made politically damaging statements. Similarly, the ability to slow down or speed up videos or edit photos has been available to the public for years.

These forgery techniques are called "Cheapfakes" or "shallow fakes." These are audio-visual (AV) manipulations created with cheaper, more accessible software, when compared to

[1] Generative Adversarial Networks, Ian J Goodfellow (2014)

deepfakes. They require less technical skill and are available on a larger scale. Their intended use however, of attempting to undermine the truth, is the same. One of the most prominent recent cheapfake videos for example, showed US Speaker of the House Nancy Pelosi appearing drunk, when in fact the video's speed was just slowed down.

The most common cheapfake techniques are:

- Relabelling
- Face swapping
- Speeding/slowing
- Footage editing
- Recontextualisation: withdrawing a certain element, a person's statement, gesture or action for example, and presenting it in a completely different context

Meta recently implemented advice from their Oversight Board, amending Meta's Manipulated Media policy to cover content that makes people appear to say and do things they did not, regardless of whether the content has been altered or generated by AI.[2] The Oversight Board flagged that non-AI-altered content is prevalent and not necessarily any less misleading. To effectively target all misleading content, it will be important to specify the harms seeking to be prevented. Through focusing on more serious risk of harm, rather than specific AI elements, the threat of cheapfakes can be simultaneously addressed. This also highlighted the value of risk assessments, which are already a crucial part of the Online Safety Act's implementation process and will be part of Ofcom's guidance this year.

The Current UK Legal Framework

The challenges presented by deepfakes concern several areas of law and policy, from copyright, data protection, online safety and intellectual property.

The UK government has introduced several initiatives to try to address deepfakes, including funding research into deepfake detection technologies and partnering with academic institutions to develop best practices for detecting and responding to deepfakes. It has funded research and development to support and spread awareness about the harms of revenge or deepfake pornography in its ENOUGH communications campaign.[3]

However, deepfakes have a range of threats and uses, and mitigating or preventative tools will need to be suitably matched. To combat the rise of AI deepfakes, UK law provides a range of possible options for those who fall victim to the technology.

Online Safety Act

The recently enacted Online Safety Act (OSA)[4], makes the sharing of intimate images and content which incites violence, including deepfakes and deepfake pornography, a criminal offence.

Criminal Justice Bill

---

[2] https://www.oversightboard.com/news/meta-makes-significant-changes-to-manipulated-media-policy/

[3] https://www.gov.uk/government/news/home-secretary-says-enough-to-violence-against-women-and-girls

[4] https://www.legislation.gov.uk/ukpga/2023/50

The Criminal Justice Bill[5], which is currently passing through the UK House of Commons and is not yet law, is seeking to criminalise the creation of deepfake pornography through an amendment. The Ministry of Justice announced that creating a sexually explicit "deepfake" image is to be made an offence. Under the legislation, anyone who creates such an image without consent will face a criminal record and an unlimited fine, regardless of whether the creator intended to share it. It will also strengthen the existing OSA offence, so if a person both creates this kind of image and then shares it, the Crown Prosecution Service could charge them with two offences, potentially leading to an increased sentence.

Privacy and Harassment

Where a deepfake depicts an individual in a private situation, for example deepfake pornography, a claim may be brought on privacy grounds. It will be irrelevant that the content is false; English law provides that if the victim has a reasonable expectation of privacy in relation to the type of information, then a remedy can be sought irrespective of its truthfulness. Where the individual has been caused alarm or distress, it may also be possible to bring a claim for harassment.

Intellectual Property

The development of deepfakes using IP such as such as a performers' images, voices, or likeness without permission is a topic that has been raised and was acknowledged by the Intellectual Property Office through public consultation. While the IPO has committed to keeping this issue under review from an IP perspective, it has stated that existing law is not clear on the impacts of AI technologies on performers, and that any intervention may not be best addressed through the UK's IP framework[6][7].

Separately, as part of its pro-innovation approach to AI regulation, the former UK Government committed to launching a call for evidence on AI-related risks to trust in information and related issues such as deepfakes.

Copyright

A victim of a non-consensual deepfake may attempt to have the deepfake removed from media platforms by seeking an injunction pursuant to a copyright infringement claim.[8] This may be difficult to establish given the number of different rights holders that may be in play and may depend on the specific images or content being used in a deepfake, as well as legal interpretation as to whether they amount to an act of 'copying' (in whole or substantial part) of the copyrighted work. Further, depending on the context and legal interpretation, deepfakes may fall under the one of the fair dealing exemptions in the Copyright, Designs and Patents Act 1998 (CDPA), for example through parody.[9] The relationship between generative AI system developers, rightsholders and copyright is an open question that is subject to debate

[5] https://bills.parliament.uk/bills/3511

[6] https://www.gov.uk/government/consultations/artificial-intelligence-and-ip-copyright-and-patents/outcome/artificial-intelligence-and-intellectual-property-copyright-and-patents-government-response-to-consultation#executive-summary

[7] https://ico.org.uk/about-the-ico/media-centre/news-and-blogs/2024/04/information-commissioner-s-office-seeks-views-on-accuracy-of-generative-ai-models/

[8] https://www.gov.uk/guidance/enforcing-your-copyright

[9] https://www.gov.uk/guidance/exceptions-to-copyright

globally, and is under consideration by the UK Government. The World Intellectual Property Organisation (WIPO) published the "Draft Issues Paper On Intellectual Property Policy And Artificial Intelligence"[10] in December 2019, which included recommendations for establishing a system of equitable remuneration for victims of deepfake misuse and addressing copyright in relation to deepfakes.

Existing consumer law frameworks can also apply in some cases, particularly where deepfakes are used to deceive or defraud individuals. Some groups have encouraged regulators to rely on existing legal frameworks to regulate, which allow for recourse mechanisms including takedown notices or legal action. For example, illegally impersonating someone for the purposes of obtaining money, goods or services is a criminal offence under the Fraud Act 2006.[11]

Defamation

Alternatively, a deepfake victim could rely on defamation legislation. The Defamation Act 2013 consolidated large parts of existing caselaw and statute in this area and established a new threshold for bringing a defamation claim.[12] Under this new threshold, a harmed individual must show that a deepfake caused or likely caused serious reputational harm, for it to be considered defamatory. The 'meaning' of any allegation in the deepfake image or video, taken as a whole, will also be material, and if the reasonable viewer is not aware of the video's falsity, it may be possible to bring a claim against the creator of the video.

Regulation in the European Union (EU) and the EU AI Act

The EU has taken a proactive approach to deepfake regulation, calling for increased research into deepfake detection and prevention, as well as regulations that would require clear labelling of artificially generated content. In 2018, the EU's Code of Practice on Disinformation[13] was initially introduced as a voluntary self-regulatory instrument to tackle the spread of deepfakes and disinformation on social media platforms. These measures are now part of the Digital Services Act[14], in force since 2022.

In the EU, deepfakes will be regulated by the AI Act[15], the world's first comprehensive AI law. The Act will not bar the use of deepfakes outright but attempts to regulate them through transparency obligations placed on the creators and providers of Generative AI. Under Article 52(3) of the Act, users of an AI system that generates deepfakes, must disclose that the content has been artificially generated or manipulated.

[10] Draft Issues Paper on Intellectual Property Policy and Artificial Intelligence (wipo.int)
[11] https://www.legislation.gov.uk/ukpga/2006/35/contents
[12] https://www.legislation.gov.uk/ukpga/2013/26/contents
[13] The 2022 Code of Practice on Disinformation | Shaping Europe's digital future (europa.eu)
[14] https://commission.europa.eu/strategy-and-policy/priorities-2019-2024/europe-fit-digital-age/digital-services-act_en
[15] https://artificialintelligenceact.eu/

Opportunities from Synthetic Media

Unleashing a new wave of creativity, synthetic media is rapidly transforming how we interact with information. From revolutionising education to fostering accessibility, this technology holds immense potential to enrich our experiences, offer more autonomy and empower innovative storytelling.

As synthetic media tools become more readily available and easy to use, they offer opportunities to eliminate monotonous tasks in content creation and to work faster and more efficiently. They also enable creative expression in new ways and enrich our experiences in gaming, virtual reality and more.

Education

Synthetic media has the potential to revolutionise education by creating more engaging and interactive learning experiences. Synthetic media can be used to animate historical photos and footage, allowing figures to come to life in the classroom. It can also provide a more affordable and accessible way to train vocational courses. For example, synthetic media is being used in medical training simulations, allowing medical professionals to practice procedures on virtual patients. 3D medical imaging solutions have also helped surgeons plan and rehearse complex surgeries. Language learning apps are similarly using this technology in an educational setting, to generate realistic conversational partners for learners. This allows students to practice their speaking and listening skills in a simulated environment.

Case Study: Synthesia.io

Synthesia.io is an AI video creation platform that empowers businesses to produce professional videos without the need for studios, cameras, or actors. Its core technology lies in synthetic media, specifically creating AI-generated human avatars that can be programmed with speech and text. However, unlike traditional deepfakes, Synthesia focuses on ethical and transparent applications.

Synthesia allows users to create personalised training videos with a virtual instructor in any language, thereby allowing companies to efficiently deliver consistent training experiences to a global workforce. It also offers opportunities for educational videos, whereby users can create engaging and informative videos without time constraints, using the platform's library of diverse avatars. The service also offers the opportunity to produce a high volume of video content for various marketing channels such as product demos and explainer videos, without the need for expensive video shoots. By democratising video creation and ensuring responsible AI, synthetic media can therefore being leveraged to foster communication, education, and training on a global scale and create positive impact.

Autonomy and Expression

Using synthetic media technology to anonymise voices and faces to protect privacy has the potential to be used in a number of positive ways. It could help activists and journalists to remain anonymous in dictatorial and oppressive regimes, helping them to report out atrocities on traditional or social media without fear of danger. This technology could further be used in instances such as witness protection, to obscure the identities of witnesses in sensitive legal cases.

Synthetic media also gives individuals new tools for self-expression and integration in the online world. For example, it offers the opportunity to create personal AI avatars. A personal digital avatar could give autonomy to individuals suffering from certain physical or mental disabilities could use synthetic avatars of themselves for online self-expression.

Deep Empathy[16], a UNICEF and MIT project, uses deep learning to learn the characteristics of Syrian neighbourhoods impacted by conflict. It then simulates how cities around the world would look amid a similar conflict. The project created synthetic war-torn images of London, Tokyo and other key cities around the world to help increase empathy for victims of a disaster region.

Business

Synthetic media images have further been used by influencers to broaden their reach and increase their audience. Using this technology, a brand can reach many customers with highly targeted and personalised messaging. A 2018 Zalando campaign featuring model Cara Delevingne across 290,000 localised ads was achieved using "deepfake technology" to produce a range of alternative shots and voice fonts.[17]

Lil Miquela, is also 'virtual influencer'- a computer generated 21 year old "robot living in LA" with over 2.5m followers on Instagram.[18] One of the first virtual influencers, created by media and creative agency Brud, Lil Miquela charges up to hundreds of thousands of dollars for any given deal and has worked with Burberry, Prada and Givenchy.

The creation of digital avatars is also being used by businesses in other sectors. Companies like Colossyan[19] also offer the Colossyan Creator tool, which allows users to create customisable AI avatars to allow for more engaging and interactive video experiences for use cases such as training and education.

Case Study: AI Foundation

Organisations such as the AI Foundation are creating personal AI for influencers and celebrities, with their consent, to engage and amplify their reach with the audience, create deeper engagement with the fans, and deliver personal experiences at scale.

In 2019, the AI Foundation launched a personal digital avatar of Deepak Chopra. Digital Deepak looks, speaks, and acts like world-renowned writer and teacher Dr. Chopra. It offers personalised advice from Dr. Chopra, allows you to ask questions on every aspect of well-being, and benefit from meditation, anywhere, at any time. Dr. Chopra believes that Digital Deepak is truly the future of well-being and exemplifies how advances in AI can unlock humanity's full potential. The AI Foundation's mission is to enable all seven billion people in the world to have their own AI that shares their values and goals.

[16] https://deepempathy.mit.edu/
[17] https://www.voguebusiness.com/companies/how-deepfakes-could-change-fashion-advertising-influencer-marketing
[18] https://www.instagram.com/lilmiquela/
[19] https://www.colossyan.com

Synthetic speech

We are not only seeing synthetic media software being used in image and text generation, but also in audio. Not only do we see generative AI computer systems creating music, but even the human voice is being generated by AI. In 2023, a health charity partnered with David Beckham to produce a video and voice campaign to help end malaria. In the "Malaria Must Die" campaign, Beckham spoke nine languages seamlessly.[20] The social campaign was an example of using deepfakes to broaden the reach of a public message. Companies like Synthesia, the company behind the Beckham video, and VOCALiD[21], a voice startup, create custom synthetic videos and voices for learning tools, brand marketing, audience engagement, customer service, and public messaging uses.

With this technology, several opportunities are presented. Game developers, filmmakers, and creators of marketing or training videos could be assisted in generating audio clips, thereby eliminating the time and cost of traditional voiceover recordings. Other companies are also developing synthetic voice technologies including Replica Studios[22], a voice AI and text to speech technology company and Lyrebird which offers voice cloning.[23]

Furthermore, voice actors could potentially benefit, as Replica Studios stated they were creating a marketplace of voices where voice actors could record and license their voices for studios to use. While studios could cut costs around hiring voice actors, the actors themselves could also make money by licensing their voices to multiple studios at the same time.

Some voice technology startups, such as HearAfter AI[24] are even creating synthetic voice as a new kind of bereavement therapy to help people remember and connect with their loved ones.

Healthcare

Synthetic media is rapidly transforming the healthcare landscape by offering innovative solutions to longstanding challenges. This technology allows for the creation of realistic yet anonymised data, unlocking a new era of medical research, improved patient care, and enhanced efficiency.

Case Study: NVIDIA

NVIDIA, in collaboration with MGH & BWH Centre for Clinical Data Science and the Mayo Clinic, is using synthetic media for CT image generation. This can be used to train medical algorithms to spot tumours more accurately, without the need for as much real sensitive patient data.

Further, it is using synthetic media to change the future of patient care. Hippocratic AI is developing task-specific Generative AI Healthcare Agents, powered by the company's safety-focused large language model (LLM) for healthcare, connected to NVIDIA Avatar Cloud Engine microservices and utilises NVIDIA NIM for low-latency inferencing and speech

[20] https://malariamustdie.com/news/david-beckham-launches-worlds-first-voice-petition-end-malaria
[21] VocaliD – Your voice AI company, bringing things that talk to life.
[22] https://www.replicastudios.com
[23] https://www.descript.com/lyrebird
[24] https://www.hereafter.ai/

recognition. These agents talk to patients on the phone to schedule appointments, conduct pre-operative outreach, perform post-discharge follow-ups and more.

Case Study: Humans.ai

Humans.ai is a company utilising synthetic media to address challenges in healthcare. Its platform leverages artificial intelligence and blockchain to create synthetic patient data.

Traditionally, a major hurdle in healthcarethese fields has been the scarcity of data available for testing. Synthetic media can generate realistic medical data, including artificial avatars that act as simulated patients in clinical trials. This innovation streamlines the process of discovering new treatments while enhancing the accuracy of disease detection.

Synthetic health data can also be used to train machine learning algorithms. These algorithms, in turn, can improve diagnostics and provide more precise treatment recommendations. The use of this technology in healthcare has the potential to facilitating groundbreaking research, accelerate the development of life-saving treatments, and improve patient outcomes.

Creativity

Synthetic media is also revolutionising various aspects of creative industries, offering a range of possibilities previously unimaginable. One notable application is the dubbing of films into different languages while preserving the original actor's performance to a considerable extent. This innovation not only enhances accessibility but also maintains the authenticity of the performance across linguistic boundaries.

Moreover, synthetic media is increasingly employed in cinema to generate realistic special effects. This capability has profound implications for storytelling and visual aesthetics, pushing the boundaries of what is achievable in filmmaking. At the Dali Museum in Florida, visitors are greeted by a deepfake Salvador Dali[25], who provides insights into his life and art, showcasing the technology's potential for immersive cultural experiences.

The integration of synthetic media into entertainment has become ubiquitous, particularly with the extensive use of computer-generated imagery (CGI) in movies. Platforms like Midjourney[26] make the use of AI for generating lifelike image assets more accessible, and so

[25] https://www.theverge.com/2019/5/10/18540953/salvador-dali-lives-deepfake-museum
[26] https://www.midjourney.com/home

empowering creators across various artistic endeavours, from graphic design to content creation to cultural preservation.

Case Study: Adobe

Responsible Innovation with Firefly

In March 2023, Adobe launched Firefly, a new family of creative generative AI models, and in May, Adobe announced Firefly's integration into Photoshop with Generative Fill. Based on Adobe's AI Ethics principles, Firefly was developed to be commercially safe, provides transparency to consumers, and respects the rights of artists and creators. Every asset produced with Firefly has an embedded Content Credential, indicating the origin and version of the file.

Content Credentials are a tamper-evident metadata, enabling creators to add extra information about themselves and their creative process, such as the creator's name, the date an image was created, what tools were used to create an image and any edits that were made. Content Credentials are free and built on the Coalition for Content Provenance and Authenticity's (C2PA's) open, global standard, meaning that the metadata can be shared across platforms and websites - beyond just Adobe products.

As well as providing transparency about the artefacts produced using generative AI, Content Credentials allow creators to attach a "Do Not Train" tag to the metadata of their work, signifying that they do not want their content to be included in AI training databases. These labels travel with their content wherever it goes and help to prevent web crawlers from using designated works. Adobe is looking to drive the adoption of an industry standard for this technology to protect creator rights and enhance their experiences using generative AI.

Case Study: Midjourney

Midjourney is a platform that uses artificial intelligence to generate high-quality synthetic media, transforming digital art creation. Launched in 2022, it employs advanced generative models to produce visual content from text descriptions, allowing users to create detailed images without extensive artistic skills.

A Discord survey of Midjourney users found that 68% of users use Midjourney for fun and 77% use it for personal purposes, highlighting its use for individual creativity. 32% of users still use the tool for utility, from creating visuals for marketing, to characters for game design. The platform's interface and range of output options makes it accessible to both amateurs and professionals. It allows anyone to experiment with styles and techniques through text descriptions, and bridge the gap between ideas and visuals.

Challenges

There are two broad categories of deepfake risk, the first concerns being deepfaked, namely having your image used in a deepfake video, whereas the second concerns being misled into believing that a deepfake is genuine. These digital manipulations can be used for a variety of malicious purposes, including spreading disinformation, impersonating individuals, and perpetrating fraud. As deepfakes become increasingly convincing and prevalent, the need to counter their negative impact is critical.

The number of deepfake videos published online has risen exponentially, with global verification platform Sumsub stating the number of deepfakes detected in Q1 2023 was 10% higher than in all of 2022, and the majority of these came from the UK. 51.1% of online misinformation also comes from manipulated images.[27]

So far, deepfakes have already been used to create realistic "revenge pornography" involving celebrities and members of the public, but increasingly they are being used to discredit politicians and business leaders or defraud companies. Wider threats deepfakes pose range from identity theft and privacy concerns to reputational damage, eroding trust in media and undermining election integrity.

Fraud

One of the most worrying risks from deepfakes is the potential to assist criminals in the commission of fraud, against both individuals and companies. The ability to look and sound like anyone, including those authorised to approve payments from the company, gives fraudsters an opportunity to exploit weak internal procedures and extract potentially vast sums of money from banks. The schemes would be more sophisticated versions of phishing and business email compromise scams, though harder to detect. Additional risks include the erosion of a brand's trust and reputation, potential market manipulation, legal and compliance concerns as well as threatening the ability to vet third parties.

Identity Fraud

Fraudsters are increasingly using deepfakes to attempt todupe identity verification systems. For example, to open illegitimate bank accounts.

Phishing Scams

Modern phishing attempts have incorporated fake video and audio messages, which are often personalised and tailored to individuals. Using deepfake technology, scammers can generate video clips of trusted figures, celebrities, or even family members, asking the recipient to undertake certain financial actions, including sharing sensitive information or transfer funds to unauthorised accounts.

Impersonation Attacks

Deepfakes can also be used to mimic corporate executives or government officials. Successful fakes can trick employees into divulging sensitive information or money. In the

[27] Sumsub launches advanced deepfakes detector | Sumsub

case of government officials, this information passed to bad actors may even be considered espionage.

Recently, a finance worker at a multinational firm was tricked into paying out $25 million to fraudsters using deepfake technology to pose as the company's chief financial officer in a video conference call, according to Hong Kong police.[28] The scam saw the worker duped into attending a video call with what he thought were several other members of staff, but all of whom were in fact deepfake recreations.

Disinformation, Misinformation and Trust

Deepfakes and advanced synthetic media serve as amplifiers of existing threats, including deceptive campaign ads, mis and disinformation. The World Economic Forum's Global Risks Report 2024[29] ranks misinformation and disinformation as the number one threat the world faces in the next two years. These issues predate the emergence of deepfakes and would continue even if AI-generated content was not present. Misinformation has long been a feature of election campaigns around the world. Photoshopped images, memes, and fake audio and video of politicians have been around for decades.

However recent events, including Slovakia's election, have seen deepfakes used to spread disinformation.[30] AI-generated audio recordings impersonating a liberal candidate circulated two days before polls, creating confusion. While fact checkers rushed to verify that the audio was fake, the candidate ultimately lost the election. While the effect of the deepfake has not been quantified, there is a risk that such fakes could have significant impacts in tight races this year.

Even today, while we do see instances such as the Keir Starmer deepfake which circulated during the 2023 Labour Party Conference[31], it is hard to point to a convincing deepfake that has misled people in a tangible or quantifiable way. Arguably, the danger lies not in deepfake videos of politicians but more so in the manipulation of content to manufacture false narratives.

The proliferation of deepfakes introduces challenges because they can deprive the public of the accurate information needed to make informed decisions, especially in elections. However, the deeper concern lies in the emergence of the "liar's dividend" — a phenomenon where the very existence of generative AI engenders an atmosphere of mistrust.

The fear is also that the sheer volume of AI-generated content could make it challenging for people to distinguish between authentic and manipulated information. Furthermore, the boom in large language models, and text-to-speech, or text-to-video software, also speed up the creation of content.

The mere existence of deepfakes could also undermine the primacy of credibility and authority of traditional social institutions, like the press, government, and academia. If the public views things not only with scrutiny, but with a default posture of doubt and disbelief, then this could be exploited by malign actors hoping to muddle reality. Malign actors could intentionally sow

[28] https://edition.cnn.com/2024/02/04/asia/deepfake-cfo-scam-hong-kong-intl-hnk/index.html
[29] Global Risks Report 2024
[30] https://www.wired.com/story/slovakias-election-deepfakes-show-ai-is-a-danger-to-democracy/
[31] https://www.politico.eu/article/uk-keir-starmer-labour-party-deepfake-ai-politics-elections/

a sense of scepticism that undermines trustworthy institutions and questions the legitimacy and authenticity of true content and media, suggesting that authentic content is an elaborate deepfake. In a climate where the political spectrum is polarised and adversarial, with 24-hour media cycles, a politician facing a critical scandal could claim something is a deepfake, thereby evading accountability and preserving their reputation.

Sophisticated actors could also synthetically reproduce a genuine event using Generative AI technology, inserting a detectable signature, to trigger a response from authentication and detection tools. This would call into question whether the real content is legitimate in the first place. This could enable malicious actors to claim the reproduced media is a deepfake and that the genuine event never occurred. Clearly, deepfakes have the potential to undermine the credibility of history.

Elections and Democracy

It is estimated that two billion people around the world will vote in national elections throughout 2024, including in the UK, US, India and 60 other countries.[32] As malicious actors seek to exploit the capabilities of deepfakes, the potential for swaying public opinion, eroding trust in democratic institutions, impacting elections and spreading misinformation about politicians rises. The question is not whether deepfakes will be maliciously deployed to spread disinformation, but rather how effectively Governments, the media, technology companies and the public will respond to their use.

Case Study: Microsoft

Content Integrity Tools to Support Global Elections

Microsoft recently announced the expansion of the private preview of their Content Integrity tools to EU political parties, campaigns and news organisations from around the world. Through this expansion it is delivering tools organisations can use to help voters understand the information they encounter online.

Microsoft built its Content Integrity tools to help organisations such as political campaigns and newsrooms send a signal that the content someone sees online is verifiably from their organisation.

When people see media with valid Content Credentials, they can be certain that the content was in fact released by the newsroom, campaign, or political party. And they can understand whether the media has been altered in any way because they can see the editing history from the time that the organisation added Content Credentials. This is made possible by leveraging the open-source industry standard published by the Coalition for Content Provenance and Authenticity (C2PA). These tools will be made available in private preview at no cost through 2024.

What's included in Microsoft Content Integrity tools?

- The content integrity tool consists of three components:

[32] https://www.weforum.org/agenda/2023/12/2024-elections-around-world/

- An easy-to-use private web application available to political campaigns, news organisations, and election officials so they can add Content Credentials to their owned, authoritative content
- A private mobile application to capture secure and authenticated photographs, video, and audio by adding Content Credentials in real-time from a smartphone, developed in partnership with Truepic
- A public website for factcheckers and any member of the public to check images, audio, or videos for the existence of Content Credentials

In February 2024, an audio deepfake emerged that mimicked the voice of US President, Joe Biden.[33] The audio clip was used in an automated telephone call targeting Democratic voters in the US State of New Hampshire. In the faked message, an AI-generated version of Biden's voice is heard urging people not to vote in the state's primary election. With a potentially divisive US election scheduled for November 2024, in which Biden looks likely to contest the presidency with Donald Trump, US authorities are drafting new laws that would ban the production and distribution of deepfakes that impersonate individuals.

We do not know how persuasive and impactful deepfakes are at changing voters' opinions or voting preferences. The efficacy of each case will likely depend on different factors, including what kind of deepfake is deployed, how it is contextualised, how it spreads, how it is timed, for example on the eve of an election or in moments of existing tension or crisis, and finally, how we respond. The Slovakia election and Biden deepfake examples show us that there are already emerging patterns and playbooks. Countering these will likely be crucial to ensuring that attempts to disrupt elections with deepfakes fail.

National Security

Criminals working on behalf of malicious states could use AI-generated deepfakes to hijack the general election, the Home Secretary said in February 2024.[34] He warned that people working on behalf of states such as Russia and Iran could generate thousands of deepfakes to manipulate the democratic process in countries such as the UK.

Since 2019, malign actors associated with nation-states, including Russia and China, have conducted influence operations leveraging GAN-generated images on their social media profiles.[35] They have used these synthetic personas to build credibility and believability to promote a localised or regional issue. This is not a singular incident, and it seems to be a common technique now in the age of influence campaigns. Social media platforms, such as Facebook, and other AI/ML research companies, such as Graphika, were able to detect these profiles and assess the images were AI-based and synthetically generated. However, detecting synthetic personas isn't always timely, and it is hard to know how many other social media profiles using GAN-generated images have not been detected. Some specific examples of synthetic content used as part of influence campaigns include:

[33] https://news.sky.com/story/fake-ai-generated-joe-biden-robocall-tells-people-in-new-hampshire-not-to-vote-13054446

[34] https://www.thetimes.com/uk/article/james-cleverly-deepfakes-threat-next-general-election-bwmjcdfpm

[35] https://www.nytimes.com/2024/05/20/world/asia/china-russia-deepfake.html

- From 2020 to 2021, social media personas with GANs-generated images criticised Belgium's position on 5G restrictions, in an apparent effort to support Chinese firms trying to sell 5G infrastructure.[36]

- In 2021, FireEye reported cyber actors used GANs-generated images in social media platforms to promote Lebanese political parties[37]

- In 2022, pro-China bot accounts promoted AI generated footage of fictitious people for state-aligned influence campaigns.[38]

Privacy and Harassment

A recent report from 2023, puts non-consensual pornographic deepfakes as now forming 98% of total deepfake content, with 99% of them targeting women.[39] From 2019 to 2023, the report found a 550% rise in total deepfake videos online. So not only is deepfake content rising exponentially, but most of it is pornographic, and the abuse of such content is gendered, overwhelmingly targeting women and girls.

The most relevant example of this is the recent publication of sexually explicit AI-generated images of Taylor Swift on X, which attracted more than 45 million views, 24,000 reposts, and hundreds of thousands of likes for nearly 17 hours before being shut down.[40] Despite being eventually shut down, the images spread and were reposted across accounts and social media platforms. This incident, and others like it, speak to the challenges in stopping deepfake pornography of real people and in controlling the damages. Given that the more content and training data fed to GANs, the better and more realistic deepfakes they can create, this poses a particular problem for public figures, as many have swathes of images online, current deepfake technologies can create particularly accurate doppelgangers.

There is also concern around how non-consensual deepfake image abuse is a threat to ordinary citizens, specifically women, given the rise in revenge pornography perpetrated by people who know their victims personally.

Legislation has begun to try to address this issue. From 31st January 2024, the Online Safety Act has made the sharing of AI-generated intimate images without consent illegal. The Act has also brought in further changes around sharing and threatening to share intimate images without consent.

[36] https://graphika.com/reports/fake-cluster-boosts-huawei
[37] https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf
[38] https://public-assets.graphika.com/reports/graphika-report-deepfake-it-till-you-make-it.pdf
[39] 2023 State Of Deepfakes: Realities, Threats, And Impact (homesecurityheroes.com)
[40] https://www.nytimes.com/2024/01/26/arts/music/taylor-swift-ai-fake-images.html

Solutions

AI-generated media is becoming increasingly difficult to detect by humans. To date, deepfakes are often easy to spot, for example when images of people feature six fingers or unusually distorted faces. Soon however, those tells will be resolved, thereby making it more difficult to detect deepfakes with the naked eye. This is where technology offers an ability to bridge the gap.

To address the issue of deepfakes, companies will need to invest in robust AI detection tools, employee training, authentication mechanisms, and collaborative efforts with industry peers. By taking proactive measures to detect, prevent, and respond to deepfake threats, businesses can maintain their reputation, protect their stakeholders, and contribute to a safer digital ecosystem. As technology advances, companies must develop and update their defences constantly, like the way businesses currently approach more conventional cybersecurity threats.

Detection and Mitigation

As technology advances, it will become increasingly difficult to identify manipulated media with the naked eye. However, there are a range of tools that can be used to help detect fake media. Every tool may vary in what they quantify as a deepfake as well, which will affect the type of media manipulation that is flagged.

techUK members are developing advanced detection tools and technologies to identify deepfakes and synthetic media across platforms, bolstering digital authenticity and credibility. Recently, a large consortium of tech companies kicked off the Deepfake Detection Challenge (DFDC)[41] to find better ways to identify manipulated content and build better detection tools. Some examples of detection technologies that have been developed in recent years and are being used by members include:

1. ***Biological signals:*** This approach tries to detect deepfakes based on imperfections in the natural changes in skin colour that arise from the flow of blood through the face.

2. ***Phoneme-viseme mismatches*:** For some words the dynamics of the mouth, viseme, are inconsistent with the pronunciation of a phoneme. Deepfake models may not correctly combine viseme and phoneme in these cases.

3. ***Facial movements:*** This approach uses correlations between facial movements and head movements to extract a characteristic movement of an individual to distinguish between real and manipulated or impersonated content.

4. ***Recurrent Convolutional Models:*** Videos consist of frames which are just a set of images. This approach looks for inconsistencies between these frames with deep learning models

5. ***Generation model specific analysis:*** Methods that rely on recognising the internal workings of how diffusion models generate image data. Examples include looking at signatures in the relationship between an image and its caption or an image and its reconstruction generated via diffusion model. These methods are less generalisable

[41] Deepfake Detection Challenge Dataset (meta.com)

to other generation techniques but are decoupled from the type of contents in an image, such as people.

Case Study: Intel

Intel is leveraging its AI expertise to tackle deepfakes by developing algorithms and software solutions that detect and mitigate manipulated content, offer responsible generative models, and media provenance. Through machine learning and advanced analytics, Intel is working to provide tools that verify media authenticity using biometrics. Last year it launched a real-time Deepfake Detector, the world's first real time deepfake detector. The detection platform utilises FakeCatcher algorithm, which analyses 'blood flow' in video pixels, alongside eye-gaze based and motion-based detection, to return results in milliseconds with 96% accuracy.

Most deep learning-based detectors look at raw data to try to find signs of inauthenticity and identify what is wrong with a video. In contrast, FakeCatcher looks for authentic clues in real videos, by assessing what makes us human—subtle "blood flow" in the pixels of a video. These blood flow signals are collected from all over the face and algorithms translate these signals into spatiotemporal maps. Then, using deep learning, they can instantly detect whether a video is real or fake.

Intel's responsible generative AI strategy is also building models with assumptions and design priors that naturally disable malicious uses; instead of patching up or filtering models after the fact. My Face My Choice, MixSyn (for multi-source image synthesis), and My Body My Choice are just a few examples in this stream.

Its media provenance vector also includes both protection of ownership and integrating provenance into the data. My Art My Choice and My Voice My Choice are two example projects in this vector, which prevent visual and audio information from being used in generative models.

Protective Shields

Every image we post online is hypothetically available for anyone to use to create a deepfake. As the latest image-making AI systems are so sophisticated, it is growing harder to prove that AI-generated content is fake. But new defensive tools are allowing people to protect their images from AI-powered exploitation, by making them look warped or distorted in AI systems.

One such tool, called PhotoGuard[42], was developed by researchers at MIT. It works like a protective shield by altering the pixels in photos in ways that are invisible to the human eye. When someone uses an AI app or image generator to manipulate an image that has been treated with PhotoGuard, the result will look unrealistic. Fawkes[43], a similar tool developed by researchers at the University of Chicago, cloaks images with hidden signals that make it harder for facial recognition software to recognise faces.

Another tool, called Nightshade[44], developed by researchers at the University of Chicago, applies an invisible layer of "poison" to images. The tool was developed to protect artists from

[42] https://www.technologyreview.com/2023/07/26/1076764/this-new-tool-could-protect-your-pictures-from-ai-manipulation/
[43] https://sandlab.cs.uchicago.edu/fawkes/
[44] https://www.technologyreview.com/2023/10/23/1082189/data-poisoning-artists-fight-generative-ai/

having their copyrighted images scraped without their consent. Yet, in theory, it could be used on other images to protect them from use by third party AI systems.

Though these defensive shields work on the latest generation of AI models, deepfake technology is advancing and it will be an arms race to ensure future versions of deepfakes cannot override these protective mechanisms. Further caveats include that these technologies do not always work on images that are already online and are harder to apply to images of celebrities.

Initiatives from companies now try to authenticate media and train moderation technology to recognise the inconsistencies that mark synthetic content. But they are in a battle to outpace deepfake creators who often discover new ways to fix defects, remove watermarks and alter metadata, creating an effective technological arms race between deepfake creators and detectors.

Cross Sector Initiatives

Case Study: AI Elections Accord

On February 16, 2024, at the Munich Security Conference (MSC), technology companies pledged to help prevent deceptive AI content from interfering with this year's global elections in which more than four billion people in over 40 countries will vote.

The "Tech Accord to Combat Deceptive Use of AI in 2024 Elections" is a set of commitments to deploy technology countering harmful AI-generated content meant to deceive voters. Signatories pledge to work collaboratively on tools to detect and address online distribution of such AI content, drive educational campaigns, and provide transparency, among other concrete steps. It also includes a broad set of principles, including the importance of tracking the origin of deceptive election-related content and the need to raise public awareness about the problem. The Accord is one important step to safeguard online communities against harmful AI content, and builds on the individual companies' ongoing work. Participating companies agreed to eight specific commitments:

1. Developing and implementing technology to mitigate risks related to Deceptive AI Election content, including open-source tools where appropriate
2. Assessing models in scope of this Accord to understand the risks they may present regarding Deceptive AI Election Content
3. Seeking to detect the distribution of this content on their platforms
4. Seeking to appropriately address this content detected on their platforms
5. Fostering cross-industry resilience to Deceptive AI Election Content
6. Providing transparency to the public regarding how the company addresses it
7. Continuing to engage with a diverse set of global civil society organisations, academics
8. Supporting efforts to foster public awareness, media literacy, and all-of-society resilience

Watermarking

Some techUK members have already been collaborating with established media outlets to consider watermarking authentic broadcast media content online.

Case Study: The Coalition for Content Provenance and Authenticity (C2PA)

In 2021, Adobe, Arm, Intel, Microsoft, and Truepic co-founded the Coalition for Content Provenance and Authenticity (C2PA) to develop technical standards for certifying the source and history of digital media. C2PA's steering committee includes a range of industry members, such as Publicis, Sony and, most recently, Google. As a mutually governed consortium which encompasses partners with diverse areas of expertise and viewpoints, C2PA is able to design careful, thoughtful architecture to help people uncover content that is authentic. C2PA unifies the efforts of the Adobe-led Content Authenticity Initiative (CAI) and Microsoft and BBC's Project Origin and oversees the development of technical standards such as Content Credentials.

Content Credentials give people more ways to find and connect with content online. They act like a digital 'nutrition' label that can show information such as the creator's name, the date an image was created, what tools were used to create an image and any edits that were made. Content Credentials help to address issues with misinformation, as they provide a technical means for trustworthy actors to be transparent about the sources, origins and creation of online media, and how and where AI was used in the process. They also allow consumers to make informed decisions about what to trust online, arming those experiencing content first hand to understand what they're viewing and where it came from.

Content Credentials are free and built on the C2PA's open standard, meaning that anyone can implement them via open-source code and toolkits - into platforms, chips, hardware, software and tools. For example, the BBC has recently introduced Content Credentials across the images and videos used within its reporting in an effort to counter disinformation when content is shared externally.

Labelling

In contrast to watermarking, another solution available is labelling via metadata. In this case, cryptographic data can be attached to an image, which is more difficult to remove, when compared to a watermark, which can be removed by compressing an image. This would operate as a label visible to users, similar to nutrition labels on food packing. It would show the extent to which a piece of content is AI manipulated or generated, and give an indication as to its authenticity, thereby allowing for content provenance at scale. When a disclosure is baked into the media itself and is very difficult to remove, it can be used as a tool to push audiences to understand how and why a piece of content was created. Though this is technically promising it is also socially challenging and will require broader change and buy in from multiple sectors and government.

Fact Checking

Fact checking will also have a role in mitigating the impact of deepfakes, especially in upcoming elections. To address the proliferation of AI-driven misinformation, several companies in the sector have introduced fact checking tools. Meta for

example, announced[45] that it has mandated the disclosure of AI-generated content in political advertisements on its platforms.

Case Study: Logically Facts

Logically Facts publishes topical, original fact-checks that align with public interest and address trending misinformation narratives. It uses data, analytics, and editorial judgement to identify and prioritise claims with significant potential for harm. Our experienced editorial team and expert fact-checkers assess the spread of individual claims online and the impacts of their dissemination.

Fact-checking plays a crucial role in tackling deepfake media, by surfacing solid evidence to disprove the content. Logically Facts has fact-checked many such videos, audio and photos circulated in countries such as the UK, Denmark, the US, Pakistan, and India.

One example came ahead of the 2024 general elections in India. A deepfake video circulated on X depicting Bollywood actor Aamir Khan criticising India's ruling Bharatiya Janata Party and expressing support for the main opposition party, the Indian National Congress. Through careful manual review, Logically Facts discovered that some parts of the audio did not sync with the actor's lip movements. Audible in the background of the video was the phrase "Satyameva Jayate", a popular TV series hosted by the actor. This prompted our team to search for their YouTube channel and confirm that the original clip had been used as a promotion for the show. An official statement from the actor labelling the viral video as "fake and completely untrue" offered further clarification. This example illustrates how fact-checkers' examination of contextual information and visual and audio clues can help to assess the authenticity of videos.

Identity Verification

Visual analysis

Deepfakes, while sophisticated, often display inconsistent facial features. AI struggles to replicate minute facial expressions, eye movements, or even the way hair and facial features interact. Algorithms that generate deepfakes can also show unnatural lighting and shadows. Visual analysis may uncover shadows inconsistent with the light source or reflections that do not align correctly.

Verification

While deepfakes can replicate voices, they might also contain unnatural intonations or subtle distortions that stand out upon close listening. Voice analysis software can help identify voice anomalies to root out deepfakes. Implementing authentication processes that layer codes or follow-up questions on top of voice commands can help ensure the request is genuine. Where files are concerned, automated document-verification systems can analyse documents for inconsistencies, such as altered fonts or layout discrepancies, that might indicate forgery.

Multi-factor authentication

[45] https://www.facebook.com/gpa/blog/political-ads-ai-disclosure-policy

Adding facial, voice, or other biometric recognition adds another hoop for a scammer to jump through even if they manage to impersonate a voice or face. Device recognition can help verify that requests are from previously authenticated or recognised devices is also an option for multi-factor authentication.

Blockchain and digital signatures

Blockchain technology promises an immutable record of all transactions. By using digital signatures and blockchain ledgers, organisations can implement provenance tracking for financial transactions to ensure authenticity and integrity. Any unauthorised or tampered transaction would lack the correct signature, flagging it for review.

Blockchain can also be harnessed to verify the authenticity of digital content. By timestamping and recording media on a blockchain, users can establish a trustworthy record of when and where the content was created. This can deter malicious actors from generating deceptive deepfakes and provide a solid foundation for authentic media verification.

Case Study: Onfido

Identity verification, crucial for maintaining the integrity of online interactions, faces a growing threat from deepfake technology. As digital services become ubiquitous, the rise of face-swap apps and Generative AI tools has facilitated deepfake creation, presenting a significant challenge for businesses and consumers alike. Onfido's Identity Fraud Report revealed a staggering 3,000% increase in deepfakes between 2022 and 2023, underscoring the urgency of the issue.

In particular, fraudsters are exploiting deepfakes and AI-generated synthetic identities to attempt bypassing document and biometric verification at customer onboarding and Know Your Customer (KYC) checks. Banks and other financial institutions have long been a primary target for opening new accounts, debit and credit card fraud, and false loan applications for example, but whatever the sector, businesses should adapt their defences or risk the consequences.

In response, AI emerges as a vital tool in the fight against deepfake fraud, leveraging algorithms to discern subtle differences between authentic and synthetic images or videos which are often imperceptible to the human eye. However, evolving fraud tactics demand constant innovation in defence strategies and enough datasets to be trained on. Onfido's Fraud Lab addresses these challenges by generating synthetic fraud examples, such as deepfakes, to train machine learning models to detect emerging fraud trends faster and more accurately. By simulating various fraud scenarios, businesses are bolstering their fraud defences, preventing substantial loses and damage.

The escalating threat of deepfake fraud necessitates a proactive approach, leveraging AI for detection and mitigation. Document and biometric verification combined with passive fraud detection signals such as geolocation and IP addresses enhances security without compromising user experience.

Media Literacy

Media literacy plays a vital role in addressing the malicious uses of deepfakes, by empowering individuals to critically evaluate information, discern between authentic and manipulated content, and make informed decisions. It acts as a proactive defence against the negative impact of deepfakes on public trust and the democratic process. Increasing the public's trust in real-time interactions and media is a long-term prospect, but a critical step to protect society from deepfakes and disinformation.

The Role of Government and Regulators

It will be critical of government to lead the policy around communicating with the public on provenance, information literacy, and what to look for when engaging with political content online. Without media literacy, it is difficult to see how there will be an effective and long-term solution to the deepfakes and disinformation threats.

Given there are legislative examples in the past which can be used as starting points, techUK and our members believe it is vital to explore this solution in greater detail. We are also keen to work closer with Ofcom's Making Sense of Media Advisory Panel, to improve the online skills, knowledge and understanding of the public when it comes to deepfakes and disinformation.

Societal education

Polices and programmes need to be set up to offer greater educational outreach to the public, addressing the issue of misinformation resistance and strengthening the public's ability to discern fact from fiction.

Experts in the field of AI/ML, including the Partnership on AI (PAI), have suggested that to improve detection, a paradigm shift should occur from focusing on detecting what is fake to bolstering what is true about the media and adding context to content. This will empower individuals to explore the authenticity of media by using context clues or metadata about where the media originated to help determine if it is real. PAI has conducted interviews that suggest individuals do not want to be told what is real or not, but rather, they want to figure it out for themselves. If individuals can legitimise the media that is being are shared, this could then improve trust in institutions.

To empower individuals to authentic media, society will have to be educated on deepfakes. Individuals may not know the true meaning of what deepfakes are or the extent of harm they can cause, but if they can be taught what to look for, they may be able to detect it on their own.

Recommendations

As we move forward, it becomes crucial for governments, industries, and society to collaboratively shape a comprehensive legal and policy framework and scale out solutions.

By actively engaging with techUK member companies at the forefront of deepfake defence, we can promote responsible AI advancement while effectively countering the threat of manipulative content, safeguarding the integrity of elections and fostering a secure digital landscape. Below we set out some recommendations on what this kind of collaboration could look like.

1. Innovation

It is essential for the government to support investment in safety tech as well as innovation in the fight against deepfakes.

Online Safety Sandbox: techUK has called for the creation of an Online Safety Sandbox[46] to support the delivery of the Online Safety Act. A sandbox is an isolated environment on a network that mimics end-user operating environments. Sandboxes are used to safely test new products and services under the oversight of a regulator.

An online safety sandbox which focuses on disinformation, misinformation and deepfakes could be used to create a vast library of synthetic media, including both real and manipulated content. This could be used to train AI models to better identify deepfakes in the real world by exposing them to a wide range of manipulation techniques.

Sandboxes are also helpful for the regulator supporting them, in this case Ofcom. It gives the regulator greater sight of what is happening in real time in the market, helping inform better regulation.

Encouraging innovation for synthetic media: while there has been a significant focus on combating deepfakes, there is a lack of dedicated funding for exploring the positive applications of synthetic media. Synthetic media should be included as a focus area within existing AI technology challenge funds. This would enable research and development (R&D) in applications across various sectors, such as education, healthcare and the creative industries.

The announcement by UKRI[47] of several million pounds for UK projects to address rapid AI advances has been a welcome step. However, the potential of synthetic media in driving positive impacts must not be overlooked. While there has been £4 million of funding through the UKRI Technology Missions Fund to support projects considering the responsible use of AI within specific contexts, further focus and funding needs to be given to explore how these technologies can improve society. For example, fostering innovation to integrate synthetic media into existing fields like EdTech and healthcare can lead to improved learning outcomes and patient care.

[46] https://www.techuk.org/resource/a-uk-tech-plan-how-the-next-government-can-use-technology-to-build-a-better-britain.html

[47] £12 million for UK projects to address rapid AI advances – UKRI

2. Standards and best practice

Creating consistent standards across tech companies and media platforms is crucial in our battle against deepfakes and disinformation. The foundation for these standards lies in content provenance: the ability to trace the origin and authenticity of digital content. While initial efforts have been made, such as the EU Code of Practice on Disinformation[48] and the AI Elections Accord[49], there remains a need for further alignment across industries and jurisdictions.

Accelerate Adoption of Content Provenance: C2PA should be supported across industry as current best practice in content provenance. This would aid the identification of authentic media and address rising threats of disinformation. By encouraging C2PA as best practice while not requiring this standard we can support the use of a tool to help tackle harms while still allowing companies to innovate and identify new content provenance tools.

Support Deepfake Detection, Labelling and Fact Checking Tools: Ofcom should promote the existing best practice that companies could use to tackle the malicious uses of deepfakes, which could include a wide range of measures such as deepfake detection, verification methods, synthetic media labelling, fact checking tools and media analysis tools.

3. Media Literacy and Collaboration

In the longer term, media literacy at both a national and global level will be integral to upskilling society and helping the public protect themselves from the dangers of deepfakes.

Promote National Media Literacy: The Government, relevant regulators and industry should collaborate to facilitate training sessions for journalists and professions likely to be targeted by deepfakes. Deepfake and disinformation focused media literacy should also form an integral part of children's education in schools. There is a particularly important role for the government to play around media literacy during elections and times of crisis.

International Cohesion: Given the global and multifaceted nature of the deepfake problem, addressing it requires a collaborative effort. Other states have begun to invest heavily in labelling and media literacy surrounding deepfakes, and it will be crucial to learn from any actions of best practice. It's also essential to find a solution that lies between legislative measures or solely relying on social media companies. By adopting a proactive stance, focusing on media literacy, and refining our strategy to combat deepfakes, the UK can position itself as a leader and innovator in best practices. This would open doors for partnerships with election watchdogs and international organisations.

[48] https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation
[49] https://securityconference.org/en/aielectionsaccord/

Glossary

Artificial Neural Networks (ANNs): A type of machine learning model inspired by the structure of the human brain.

Biometrics: Measurable biological characteristics that can be used for identification, such as facial features or voice patterns.

Cheapfake (also Shallowfake): A manipulation of audio visual content performed with a cheap, accessible and easy to use software, or without any software at all

Content Provenance: The origin and history of a piece of content, which can help establish its authenticity.

Cryptographic hashes: Generates a unique string from content as a digital signature to verify authenticity.

Deepfakes: a specific subset of synthetic media that focus on manipulating or altering visual or auditory information, to create convincing fake content, ranging from images, audio and video. A common example of deepfake use is videos that replace one person's face with another.

Deep Learning: A subset of ML that uses ANNs with multiple layers to learn complex patterns from data.

Disinformation: false information which is deliberately intended to mislead—intentionally misstating the facts.

GAN: A generative adversarial network (GAN) is a type of artificial intelligence algorithm used to generate new data samples from a training dataset. It is composed of two neural networks, a generator and a discriminator, that compete with each other in a zero-sum game. The generator creates new data samples that are similar to the training data, while the discriminator tries to distinguish between the generated samples and the real training data.

Generative AI: A type of artificial intelligence used to create new content, including images, videos, and text.

Machine Learning (ML): A branch of AI that allows computers to learn and improve without explicit programming.

Misinformation: false or inaccurate information—getting the facts wrong.

Synthetic media: any type of content whether video, image, text or voice that has been partially or fully generated using artificial intelligence or machine learning. Types of synthetic media can range from AI written music, text generation such as OpenAI's ChatGPT, computer generated imagery (CGI), virtual reality (VR), augmented reality (AR) and voice synthesis.

Watermarking: Embeds invisible markers in digital media to authenticate ownership and reveal unauthorised usage.