

Generative AI and deepfakes: a human rights approach to tackling harmful content

Felipe Romero Moreno

To cite this article: Felipe Romero Moreno (29 Mar 2024): Generative AI and deepfakes: a human rights approach to tackling harmful content, International Review of Law, Computers & Technology, DOI: [10.1080/13600869.2024.2324540](https://doi.org/10.1080/13600869.2024.2324540)

To link to this article: <https://doi.org/10.1080/13600869.2024.2324540>



© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group



Published online: 29 Mar 2024.



Submit your article to this journal [↗](#)



Article views: 200



View related articles [↗](#)



View Crossmark data [↗](#)

Generative AI and deepfakes: a human rights approach to tackling harmful content

Felipe Romero Moreno

Hertfordshire Schools of Law and Education, University of Hertfordshire, Hatfield, UK

ABSTRACT

The EU's Artificial Intelligence Act (AIA) introduces necessary deepfake regulations. However, these could infringe on the rights of AI providers and deployers or users, potentially conflicting with privacy and free expression under Articles 8 and 10 of the European Convention on Human Rights, and the General Data Protection Regulation (EU) 2016/679 (GDPR). This paper critically examines how an unmodified AIA could enable voter manipulation, blackmail, and the generation of sexual abusive content, facilitating misinformation and potentially harming millions, both emotionally and financially. Through analysis of the AIA's provisions, GDPR's regulations, relevant case law, and academic literature, the paper identifies risks for both AI providers and users. While the AIA's yearly review cycle is important, the immediacy of these threats demands swifter action. This paper proposes two key amendments: 1) mandate structured synthetic data for deepfake detection, and 2) classify AI intended for malicious deepfakes as 'high-risk'. These amendments, alongside clear definitions and robust safeguards would ensure effective deepfake regulation while protecting fundamental rights. The paper urges policymakers to adopt these amendments during the next review cycle to protect democracy, individual safety, and children. Only then will the AIA fully achieve its aims while safeguarding the freedoms it seeks to uphold.



KEYWORDS

Deepfake regulation; human rights; generative AI

1. Setting the stage: deepfakes, the EU AI Act, and the road ahead

Born on Reddit in 2017, 'deepfakes' – a combination of 'deep learning' and 'fake' – started with AI-swapped celebrity faces in videos. Initially confined to pornographic content, the practice sparked ethical concerns when the code was shared, enabling widespread creation (Thi Nguyen et al. 2022, 1). Research reveals a staggering 550% increase in AI-manipulated photos between 2019 and 2023, demonstrating their alarming accessibility and potential for misuse (Home Security Heroes. 2023).

The challenge is that the simplicity of creating realistic deepfakes and fake media environments is no longer limited to experts. Not only can deepfake videos be created,

CONTACT Felipe Romero Moreno  f.romero-moreno@herts.ac.uk  Hertfordshire Schools of Law and Education, University of Hertfordshire, de Havilland Campus, Hatfield AL10 9EU, UK

© 2024 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group
This is an Open Access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/4.0/>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

but equally fake accounts are used to amplify them: X accounts which post links to the video, accounts that comment on it, sites that host it and create misleading information, and Instagram accounts that generate memes using it (Van der Sloot and Wagensveld 2022, 2). For instance, attacks targeted Khan and Starmer, UK figures, in 2023; manipulated audio was used to smear their reputations, with a fabricated clip of Khan downplaying Armistice Day circulating among extremists (The Guardian News, November 10, 2023), and deepfakes of Starmer depicting him mistreating staff surfacing during the Labour Party conference (Sky News, October 9, 2023). Similarly, in the US in January 2024, explicit deepfakes of Taylor Swift garnered millions of views, including an astonishing 47 million on platform X, before takedowns. This ignited urgent calls for laws criminalising such harmful and invasive content (BBC News, January 27, 2024).

The ease with which manipulations can be created, particularly using generative AI tools, poses an existential threat to the integrity of information and trust in public discourse. As highlighted in the *Clarkson v OpenAI* case in the US California Northern District Court, these technologies were not only creating realistic ‘deepfakes’ but were also being actively used to disseminate misinformation, extort victims, and access classified information, making it increasingly challenging for both humans and AI to identify and verify information.¹

Generative AI’s potential is undeniable, powering chatbots like Gemini or ChatGPT and artistic creations like Sora. This potential, however, comes with a dark side: the ease of creating harmful deepfakes. Platforms like CivitAI, built on these models, have been linked to the lucrative proliferation of deeply unsettling deepfakes, also targeting vulnerable groups like children (BBC News, February 15, 2024). Despite platform policies against such misuse, loopholes and ineffective enforcement leave victims exposed.

To address this issue, the European Union’s AI Act, which was officially approved by the European Parliament on March 13, 2024 (hereinafter the AIA), defines ‘deep fake’ in Article 3 (60) as synthetic or manipulated image, audio, or video content, which would deceptively seem to be truthful or authentic, and that resembles existing individuals, places, objects or other events or entities. The EU AI Act categorises AI systems based on their potential risk, imposing stricter regulations for higher risks. According to the literature, the EU AI Act addresses the ‘deepfakes problem’ through three key elements: (1) its definition (Article 3(60)); (2) transparency obligations for AI providers and deployers also known as deepfake creators (Article 50); and (3) Recitals 132 to 137 (Labuz 2023, 8). Although treated as exceptions with ‘limited’ or ‘specific’ risk classifications, literature brings to light that deepfakes create a separate category (Edwards 2022, 7; Kop 2021, 5; Veale and Borgesius 2021, 108; Wahlster and Winterhalter 2022, 68), thus emphasising the lack of research exploring their compatibility with the European Convention on Human Rights (ECHR) and the General Data Protection Regulation (EU) 2016/679 (GDPR). Building on the author’s previous work (Romero-Moreno 2019; 2020), this research examines this compatibility, critically assessing whether implementing deepfake provisions without safeguards could violate the right of AI providers and users to privacy and freedom of expression under Articles 8 and 10 of the ECHR, and the GDPR. Notably, it proposes two amendments to address a literature gap. First, mandate structured synthetic data for deepfake detection to enhance security and protect fundamental rights. Second, classify AI intended for deepfake electoral disinformation, extortion, and AI sexual abuse content as ‘high-risk’ due to its significant potential for harm and violation of rights.

The paper proposes clear definitions, transparent oversight, and robust safeguards to effectively regulate deepfakes while protecting fundamental rights. This approach is urged to safeguard democracy, individual safety, and children, ensuring the AIA achieves its goals without compromising freedoms.

2. Navigating the double-edged sword: deepfakes' potential and peril

2.1. Unlocking the power of good: exploring beneficial applications of deepfakes

Deepfakes possess a diverse range of applications beyond malicious intent. They are employed in various industries, including visual effects, entertainment, and education. For instance, deepfakes can be used to create realistic special effects in movies, personalise digital avatars, or generate synthetic voices for individuals who have lost theirs (Forbes Magazine, July 22, 2019). Additionally, they find applications in gaming, virtual reality, photography, and entertainment, allowing people to virtually try on clothes, experience foreign films with dubbed voices, or engage with historical figures through reanimation (Mirsky and Lee 2020, 1). However, despite these positive applications, the potential for misuse of deepfakes is a growing concern.

2.2. Revealing the dangers: exploring the malicious use of deepfakes

Given the amount of data available online for public figures such as celebrities and politicians, deepfakes are often used to target these individuals. This poses a significant threat to international security, as generative AI can also be used to create deepfakes of world leaders giving speeches that are not genuine, or fabricated satellite images that include objects that do not really exist (Defense One 2019). Moreover, even without being shared on social media, intelligence services could potentially use deepfakes to target presidential advisors in an attempt to control global decision-making (CFR 2018).

The EU Charter of Fundamental Rights (Charter) safeguards the right to vote in Article 39. Recognising this, the EU AI Act classifies AI that can influence elections, referendums, or individual voting behaviour as high-risk (Recital 62). Similarly, the case *Clarkson v OpenAI* highlighted how deepfakes could also meddle in elections, sow distrust, and jeopardise public discourse.²

It is crucial to remember that deepfakes can inflict harm on anyone, regardless of their status or position. Along with the FBI, *Clarkson v OpenAI* sounded the alarm on the rising threat of 'sextortion' using generative AI and public domain images and videos of ordinary individuals. These deepfakes were created through social media to fabricate pornography. Alarmingly, this content involved not only non-consenting adults, but also kids. It was then shared widely on pornographic sites and public forums to harass the victim, inflicting significant psychological harm.³ Potentially worse, malevolent actors extorted money, sometimes requesting real-life instances of the victim performing the actions portrayed in the fabricated sexually-explicit material, by threatening to distribute the content to the victim's family, friends, and contacts.⁴

Furthermore, the case of *Clarkson v OpenAI* highlighted how the Dall-E model, due to its lower technical barrier compared to previous systems, facilitated the widespread production of abusive sexual imagery.⁵ Dall-E was trained on a massive dataset of images collected without knowledge or permission, many of which displayed real kids. Disturbingly,

sometimes real images of existing sexual abuse were deployed to train the model, creating further explicit content of previously abused kids, thereby re-traumatizing them and exacerbating their pain.⁶

Deepfakes can also be used to lure individuals into investing or giving away their hard-earned cash. For instance, a video of MoneySavingExpert Martin Lewis was widely shared on social media, using generative AI to create a realistic-looking image and voice of the journalist promoting a fake Elon Musk investment opportunity in Quantum AI. Unfortunately, the opportunity was a scam, not a legitimate investment (MoneySavingExpert News, July 7, 2023). Additionally, the popularity of AI-generated images on dating apps is increasingly growing, not only in catfishing expeditions, but also as individuals use generative AI such as Midjourney to enhance their appearance and prey on people's vulnerabilities (Los Angeles Times, May 11, 2023).

2.3. Unmasking deepfakes: exploring techniques for detection and verification

In addition to deepfake creation, there are methods for deepfake detection, one example of a well-known system is Sensity, which recognises AI-manipulated media and synthesis techniques such as AI-created faces incorporated into social media profiles, and realistic video face swaps. Sensity is trained on millions of gan-generated images to identify imperfections and small details of AI-created images (Sensity 2023). Moreover, another popular system is Intel's FakeCatcher, which, using Photoplethysmography, analyses the movement of blood vessels in a video. The colour of veins changes as the heart pumps blood through them. These 'blood flow' signals are extracted from the face and then, FakeCatcher can reliably identify real and fake videos (Intel 2022).

It is noteworthy that, while Sensity claims that it can identify realistic full bodies and faces generated using AI models like Dall-E with 98.8% accuracy (Sensity 2023), Intel asserts that its FakeCatcher technology is the first real-time system, with 96% precision (Intel 2022). In this context, the EU Court of Justice (CJEU) Advocate General (AG) opinion in *Poland v Council and Parliament* stressed that if filtering content would inevitably lead to a significant number of 'false positives,' rendering it ineffective, such measures should be excluded.⁷ Furthermore, in *UPC Telekabel Wien* the CJEU held that, to strike a fair balance, it was crucial under Article 16 of the EU Charter, to allow companies to choose the measures they will take, considering their capabilities and resources.⁸ However, the difficulty is that *Clarkson v OpenAI* also warned about the overfitting problem, where a torrent of AI-generated child abuse images confused the monitoring system since it was designed only to filter and block familiar images of abuse, but worryingly, not recognise newly created ones.⁹

3. Deepfakes and the EU Artificial Intelligence Act

3.1. AI transparency: shared obligations for providers and deployers

3.1.1. Ensuring AI content is clearly labelled and identifiable through provider responsibility

To address deepfakes, the EU's AI Act promotes transparency with Article 50(2). It requires providers of general-purpose AI tools to tag AI-generated content and identify manipulations, enabling users to better understand the information. However, this does not apply

to standard editing tasks like minor corrections or where authorised for law enforcement activities like crime detection or prosecution.

The AIA, particularly Recital 133, acknowledges the need for flexibility to accommodate various content formats, cutting-edge technologies, and AI functionalities. This ensures efficient compliance for providers, especially those dealing with diverse content and evolving technologies. Recital 133 further emphasises the importance of accurate, compatible, and effective tools for tagging and identification including technologies like watermarks, metadata tags, fingerprints, or security features to trace content origin and prove authenticity.

Interestingly, Recital 133 seems to indirectly reference the Coalition for Content Provenance and Authenticity (C2PA), an industry alliance championing content transparency. Notably, C2PA's Content Credentials 'CR' icon acts as a digital X-ray providing a detailed breakdown of the content, similar to a food label, revealing details like creator information, creation date, tools used (including generative AI), and any edits it has undergone (C2PA 2024).

3.1.2. Demanding transparency about AI in content: a legal requirement for creators

The EU AI Act takes aim at the growing concern of deepfakes, requiring their creators to inform the public about the artificial nature of their work. Article 50(4) mandates that those using AI to create deepfakes, be they creators, artists, or anyone else, must disclose this fact to the public. Recital 134 emphasises this need for transparency, stating that such disclosure should be readily apparent through clear labels and mentions of the AI origin.

However, Article 50(4) of the AIA acknowledges two key exemptions. Firstly, when investigating crimes or gathering evidence, law enforcement agencies are exempt from disclosing the use of deepfakes, prioritising legal objectives over immediate transparency. Secondly, to safeguard freedom of expression and artistic creation (protected by EU Charter Articles 11 and 13), demonstrably creative, fictional, satirical, or artistic deepfakes like movies or art exhibits have a limited disclosure obligation. Here, AI origin disclosure should be subtle, ensuring transparency without disrupting artistic expression (Recital 134).

Yet, humour and parody, crucial for social commentary, raise concerns about loopholes. Humour exemptions in free speech could create a legal grey area for harmful narratives disguised as satire (Labuz 2023, 24). Claiming 'just a joke' might enable hate speech and normalise extremism, potentially weaponizing humour itself (Ajder and Glick 2021).

The AIA also mandates disclosure of AI-generated public interest text, excluding lawful crime-fighting or human-edited content (Article 50(4)). While specific application requires careful consideration, the criteria established in the European Court of Human Rights (ECtHR) *Halet v Luxembourg*¹⁰ ruling (seriousness, contribution to public debate, and novelty), could potentially be applied to determine information disclosure requirements for AI-generated content or even labelling AI-written text as 'deepfake.'

3.1.3. Impact of the EU DSA on provider and deployer responsibilities

Recital 136 of the AI Act emphasises the crucial role of detection and disclosure of manipulated content by both providers and deployers of 'certain' AI systems. This is particularly relevant for very large platforms and search engines, who must actively address

'systemic risks' stemming from deepfakes and their potential to harm democratic processes, open discourse, and elections through disinformation. This proactive approach is essential for effectively implementing the EU Digital Services Act (DSA).

Recital 136 further clarifies that mandatory AI content labels do not replace existing obligations for hosting providers to handle illegal content notifications under the EU DSA (Article 16(6) and 16(1)). Nor does meeting the AIA's transparency requirements automatically confer legal status on the system or its outputs. Additional transparency obligations under existing legal frameworks remain in place (Recital 137).

The effectiveness of the AIA in tackling deepfakes remains questionable. A key concern involves its ambiguity regarding deepfake classification. While it requires disclosure of AI-generated content, the AIA avoids explicitly designating deceptive deepfakes as high-risk, a decision criticised by the German Bundesrat. This omission raises concerns about fragmented implementation and varying levels of protection across Member States, potentially undermining public trust and creating a regulatory patchwork (German Federal Council 2021).

3.1.4. Enforcement strategies: implementing acts, codes of practice and practical guidelines

Recognising the need for practical support despite the AI Act's mandatory deepfake labelling and detection requirements (Article 50), Article 50(7) and Recital 135 of the AIA empower a two-pronged approach led by the AI Office and the Commission.

Firstly, the AI Office facilitates the drafting of EU-level voluntary codes of practice under Article 50(7). These codes aim to simplify compliance with deepfake detection and labelling obligations for providers and deployers of manipulated and AI-generated content. The Commission approves these codes as guidance but can adopt stricter rules through implementing 'binding' acts if deemed insufficient based on a specific procedure.

Secondly, the Commission encourages the development of codes addressing broader challenges beyond just deepfake labelling and detection. These codes may target issues like making detection tools more accessible, fostering collaboration among stakeholders, and improving content provenance and authenticity transparency to empower public identification of deepfakes.

Further, Article 96(1)(d) AIA mandate the Commission to develop practical guidelines. These guidelines will specifically support the implementation of deepfake transparency obligations for providers and deployers outlined in Article 50.

However, the EU Parliament criticises instruments such as the Act's suggested codes of practice and practical guidelines as they lack transparency, accountability, and legal clarity, potentially bypassing crucial oversight and raising concerns about harm (EP 2007, 1).

3.2. Analysing AI risk: understanding the EU AI Act's classification system

The EU AI Act implements a risk-based approach to regulate deepfakes, tailoring obligations for providers and users based on the potential harm an AI system poses. However, the classification system defining 'unacceptable risk,' 'high risk,' and 'specific risk' regarding deepfakes lacks clarity, creating challenges for implementation. This paper delves into this issue, analysing the ambiguity and its potential implications.

3.2.1. Unacceptable risk: prohibited AI practices under the EU AI Act

Under the AI Act, ‘unacceptable risk’ encompasses a limited set of AI applications that pose a grave threat to fundamental rights. For example, the AIA explicitly prohibits the use of AI to manipulate cognitive behaviour or vulnerable communities, profile individuals based on their conduct, socioeconomic status, or personal qualities, or remotely identify them through facial recognition (Article 5, Chapter II).

Article 5(1)(a) of the AIA outlaws the marketing, deployment, or utilisation of any AI system that employs subliminal manipulation or deceptive techniques to inflict significant harm upon individuals or groups. This harm is achieved through impairing their ability to make informed decisions and causing them to make detrimental choices they would not have made otherwise.

While Article 5 bans harmful AI practices, it fails to address the rapidly growing threat of deepfake extortion schemes. This dangerous tactic, which, as exemplified in *Clarkson v OpenAI*, weaponizes AI-generated fabrications to coerce victims into performing sexually explicit acts under the chilling threat of public exposure. Deepfake extortion inflicts immense psychological distress, financial burdens, and irreparable reputational damage,¹¹ exposing a glaring gap in the current legal framework. Article 5(1)(a)’s clear focus on preventing manipulative practices directly applies to this scenario. Its absence of specific deepfake regulations is therefore an alarming oversight, making its silence on the matter even more concerning (Leiser 2023, 1–20).

Building on the thought-provoking and persuasive Keese-Leiser framework, one can vividly grasp the multifaceted threat of deepfake extortion to freedom of thought, a threat deeply rooted in hidden tactics, exploited vulnerabilities, manipulation of core beliefs and autonomy, and repeated or sustained exposure (Keese and Leiser 2024, 15). Deepfake extortion’s emphasis on hidden tactics and exploited vulnerabilities shines a light on how these attacks manipulate deeply held beliefs and core aspects of autonomy through repeated or sustained exposure.

Examining deepfake extortion through the lens of Article 5(1)(a) reveals its disturbing connection to hidden tactics. These tactics include harnessing social media data and AI to create deepfakes of unsuspecting individuals, often targeting vulnerable groups like women and children (*Clarkson v OpenAI*).¹² This predatory practice can wreak havoc on victims’ lives, causing long-term psychological harm (e.g. anxiety, panic, depression, PTSD) (Healthnews, November 16, 2023), shattering reputations, leading to job loss and social isolation (The Guardian News, October 28, 2023), and even causing financial issues through extortion (BBC News, September 24, 2023). In extreme cases, it can even lead to loss of life (BBC News, January 26, 2024).

However, a critical gap remains in the AIA’s framework. While Article 5(1)(a) clearly prohibits manipulative practices, it lacks regulations specifically addressing the emerging and nuanced threat of deepfake extortion. This silence, as Equality Now aptly observes, leaves legal interpretations uncertain, effectively creating a legal vacuum where victims feel powerless, unheard, and with limited options for legal recourse (Equality Now 2024, 4). In response to this gap, the following section critically analyses the legal grounds for applying the EU AI Act’s ‘high-risk’ provisions to deepfakes. It builds a case for classifying deepfake extortion, AI-generated child sexual abuse content, and electoral misinformation as high-risk AI, urgently requiring regulation to safeguard fundamental rights.

3.2.2. Navigating the gray area: regulating high-risk deepfakes under the EU framework

AI systems that negatively impact safety or the fundamental rights enshrined in the EU Charter are considered 'high-risk' AI. These are a small but significant number of AI systems which are subject to conformity assessments before market entry, ongoing monitoring, and EU database registration (Art. 6(4), Art. 49(2), Art. 71).

To protect voting rights enshrined in Article 39 of the Charter, the AIA classifies certain AI used to influence elections or manipulate behaviour as high-risk AI (Recital 62 and Annex III 8(b)) but excludes AI for campaign logistics with limited user interaction, like managing databases.

Deepfakes warrant high-risk classification due to their potential for electoral content manipulation, including spreading misinformation or impersonating candidates. This targets harmful uses while allowing legitimate campaign applications like creative video editing or content creation. The UK National Cyber Security Centre warns that generative AI models pose a significant threat to elections due to their ability to produce realistic deepfakes of politicians spreading false information, potentially reaching millions, and swaying public opinion (NCSC 2023). This concern aligns with the issues raised in *Clarkson v OpenAI*,¹³ and is echoed by the UN, which highlights the potential of deepfakes to disrupt democratic processes, exacerbate social divisions, and even threaten peace in conflict zones (United Nations 2023, 15, 18). Notably, just before the 2023 election in Slovakia, a manipulated audio clip emerged purporting to show Michal Šimečka, leading candidate for the liberal Progressive Slovakia party, discussing election rigging. While the clip's authenticity and impact on the outcome remain under investigation, it coincided with his defeat by the pro-Moscow Smer-SSD party (BBC News, December 21, 2023). This is why AI tools for misinformation deserve 'high-risk' classification. Their ability to deceive voters and manipulate public opinion poses a significant threat to democratic processes.

Google's groundbreaking initiative to disclose synthetic election ad content like deepfake electoral disinformation (Google 2023), while commendable, cannot fully secure fair elections. Alongside public education, legislation for deepfake tools, and industry standards for platforms are crucial to combat deepfake election disinformation and protect democracy.

Article 7(1) of the AIA allows the Commission to expand the list of high-risk AI systems in two ways. Firstly, by adding new areas or applications if the AI system is designed for uses listed in points 1–8 of Annex III and presents a comparable or higher level of risk to health, safety, or fundamental rights compared to current high-risk systems. Secondly, under Article 7(2), the Commission can also modify the classification of existing high-risk systems based on several factors:

1. Harm history: This includes past incidents, documented concerns, and potential for future harm.
2. Impact scope: This considers the intensity, number of people affected, and potential for discrimination against vulnerable groups.
3. Existing regulations: This factor assesses whether EU laws are sufficient to manage the risks and ensure redress.

In 2023, the potential of deepfakes to cause real-world harm materialised with alarming clarity, mirroring the warnings issued in *Clarkson v OpenAI*.¹⁴ Over 30 New Jersey high school students became victims of non-consensual deepfakes generated from their own photos, exposing the vulnerability of young individuals in the digital age (NBC News, January 17, 2024). This was not an isolated incident. ABC News reported a case in Almedralejo, Spain, where deepfakes of over 30 underage females (aged 12–14) were shared on messaging apps, used for extortion through threats of public humiliation (ABC News, October 2, 2023). These harrowing instances tragically confirmed the concerns raised in *Clarkson v OpenAI*.¹⁵ Moreover, a 2023 Stanford study revealed a deeply troubling reality: popular AI image generators like Stable Diffusion were trained on thousands of explicit images of minors, raising serious concerns about their potential to exacerbate child sexual abuse (Thiel 2023, 11). The issue extends beyond isolated cases. Reports from the Internet Watch Foundation (IWF 2023) and BBC in 2023 documented individuals exploiting similar tools to create and sell ‘synthetic’ content, primarily child sexual abuse material, on platforms like Patreon (BBC News, June 28, 2023).

Moreover, a chilling report by cybersecurity firm Home Security Heroes unveils a surge in deepfakes, exposing a growing societal threat. In 2023, identified deepfake adult videos skyrocketed by a staggering 550%, reaching a concerning 95,820 (Home Security Heroes. 2023). This exponential growth extends beyond creation, with views on the seven of the top ten adult websites hosting deepfakes amassing a massive 303,640,207 – a testament to their alarming popularity. Perhaps most troubling, 20% of viewers now prioritise deepfakes over traditional adult content, highlighting a concerning shift in media consumption (Home Security Heroes. 2023). The report paints a disturbing picture of the human cost. Non-consensual deepfakes, once primarily targeting celebrities, surged by 464% in 2023, becoming an insidious weapon against ordinary individuals (Home Security Heroes. 2023). As previously discussed, victims of deepfakes can suffer devastating consequences, including lasting psychological trauma, reputational damage, social isolation, financial harm, and even in extreme cases, loss of life. This crisis disproportionately impacts vulnerable groups, particularly women and children. The report reveals that while 48% of men have encountered deepfakes, a staggering 99% of this content features women, perpetuating harmful stereotypes and reinforcing existing societal prejudices. Deepfakes can also weaponize nationality, with malicious actors exploiting geopolitical tensions and cultural biases to fuel targeted discrimination (Home Security Heroes. 2023).

Additionally, existing EU laws have limitations in effectively mitigating risks and guaranteeing redress for online gender-based violence, particularly regarding deepfakes used for extortion and AI-powered sexual abuse. These limitations include challenges in managing risks, ensuring victim compensation, and overcoming legal loopholes. Consequently, victims face increased vulnerability due to inconsistent legal protections across jurisdictions and limited, untested options for seeking justice through existing copyright, privacy, and data protection laws (Equality Now 2024, 4, 8). While EU legislation exists aiming to protect against online violence, like the Directive on combating violence against women and domestic violence, significant gaps remain in safeguarding victims of online gender-based violence, especially concerning deepfakes used for sexual abuse. Notably, the EU DSA’s focus on illegal content, and Article 8 ECHR’s protection of image rights are valuable, but they fall short in effectively addressing the unique challenges posed by deepfakes (Equality Now 2024, 5).

The yearly review of the EU AI Act (Art. 112(1)) is insufficient to address imminent threats posed by deepfake AI, such as voter manipulation, extortion, and child abuse content generation. This inaction jeopardises democracy, individual safety, and children. The Commission is urged to swiftly reclassify these practices as ‘high-risk’ AI (Art. 7(2)) and implement stricter controls. Swift action is imperative to mitigate these pressing threats.

3.2.3. Beyond high-risk: how the EU AI Act addresses deepfake specific risk AI systems

Recital 132 of the AIA identifies specific concerns regarding ‘certain’ AI systems that interact with individuals or generate content, emphasising their potential ‘specific risks’ for deception and impersonation (e.g. deepfakes), regardless of their high-risk designation. Consequently, the AIA mandates specific transparency requirements for such systems under certain circumstances, without infringing upon existing obligations for high-risk AI. However, targeted exemptions exist for law enforcement activities like crime detection and prosecution.

Recital 132 further stipulates that individuals must be notified when interacting with an AI system. While not explicitly stated in the AIA, the emphasis on individual notification as ‘natural persons’ aligns with established legal principles like the ‘average consumer’ standard.¹⁶ The notification applies when the interaction’s nature or context would not be immediately clear to a reasonable person that they are interacting with an AI.

To ensure effective user notification, Recital 132 highlights the importance of considering age-related or disabled groups (e.g. children or individuals with cognitive impairments) when the AI system directly interacts with them.

Recital 132 also emphasises informing individuals when AI systems process their biometric data and infer sensitive elements like emotions, intentions, or categories like sex, age, or ethnicity. Crucially, these notifications must be accessible to people with disabilities. While a recent \$25 million fraud case involving deepfakes impersonating senior management underscores the dangers of AI-generated biometrics (Biometric Update 2024), exploring it falls outside this paper’s scope.

However, a closer look at CJEU caselaw reveals missed opportunities to strengthen notification requirements. Building upon their evolving focus on clear and specific notification, the AIA could have further refined requirements, particularly for interactions with potentially deceptive and often highly-risky AI like deepfakes.

For instance, a harmful deepfake might involve a seemingly genuine news video featuring a politician making false or inflammatory statements. The AIA could have countered such threats by mandating ‘sufficiently precise’ and ‘adequately substantiated’ notifications on hosting platforms (meeting the EU DSA Article 6 knowledge threshold) for deepfakes used in electoral disinformation, extortion, or AI-generated child abuse material. These clear warnings would empower operators to make context-aware judgments (like the *L’Oréal v eBay*¹⁷ case) and enable platforms to swiftly remove harmful content without extensive legal analysis, while respecting freedom of expression (as in *YouTube v Cylando*).¹⁸

Furthermore, while safety notices require further research, the AIA could have significantly bolstered user protection by mandating differentiated warning labels tailored to specific content categories. Drawing inspiration from the CJEU *Wirtschaftsakademie*¹⁹ case emphasising shared platform and creator data accountability, this collaborative

approach could demand nuanced risk communication through targeted warnings (e.g. high-risk, moderate-risk, and low-risk categories). This would further safeguard vulnerable individuals, particularly children and those with cognitive impairments, who are more susceptible to online manipulation.

Ultimately, transparency safeguards user rights in two key ways. Firstly, as *Tele2/Watson* emphasised, it empowers users to seek legal remedies against potentially harmful AI content like deepfakes.²⁰ Secondly, it aligns with the CJEU's *SCHUFA (Scoring)* position on automated decision-making (GDPR Article 22(1)), ensuring clear and meaningful information for users to understand the logic, significance, and potential consequences of AI decisions, thereby empowering legal remedies, and protecting against discriminatory consequences.²¹

4. Deepfakes and the GDPR: understanding the regulatory landscape

To uphold fundamental privacy and data protection rights (EU Charter, Articles 7 and 8), Recital 69 AIA requires data minimisation and privacy by design/default, promoting responsible AI development throughout its lifecycle. Providers should comply by using methods like anonymization, encryption, and privacy-enhancing technologies like federated learning, which enables training on decentralised data.

However, deepfakes pose unique challenges. The GDPR defines personal data as 'any information relating to an identified or identifiable natural person' (Article 4(1)). Creating a deepfake typically involves using data linked to a specific person, like voice recordings, photos, or videos. When a deepfake portrays a real individual, it clearly falls under the GDPR's purview due to its identifiable nature (EP 2021, 38).

Yet, a crucial question remains: if a deepfake is proven to be fictional, does the data used to create it still qualify as personal data under the GDPR? Generating a deepfake of a fictional person avoids targeting any specific individual. While the final deepfake might not qualify as personal data under Article 4(1) GDPR, the process itself raises concerns.

The key lies in the input data. Even if randomly generated, it may still subtly reflect characteristics of real individuals used to train the tool. These digital fingerprints could potentially be exploited to re-identify those individuals, especially if their features were unique. Therefore, even when creating deepfakes of fictional people, using input data derived from real individuals might still trigger GDPR compliance requirements due to the potential for re-identification (ICO 2023, 85).

The GDPR's broad definition of 'processing' pervasively regulates deepfakes, encompassing every stage from data collection to distribution. This stems from their reliance on personal data, triggering GDPR application. Even developers not actively creating deepfakes become subject to the regulation if they use personal data for algorithm training. Similarly, creators and distributors face scrutiny due to their use of such data in crafting and sharing deepfakes. Essentially, the GDPR makes no distinction between deepfake creation stages; if personal data is involved, meticulous compliance is mandatory (EP 2021, 39).

4.1. Deepfakes, personal data, and article 6 GDPR

Under the GDPR, personal data processing must always have a legal basis. Of the six justifications specified in Article 6, arguably only consent and legitimate interests for creating

deepfakes may be relevant (EP 2021, 39). However, obtaining individual consent (informed, freely given, specific, and unambiguous as per GDPR and CJEU rulings in *Planet49*²² and *Orange România*)²³ presents the first possible way to legally use personal data for deepfakes. This means individuals in both the unmanipulated and manipulated content must have actively agreed to the personal data processing and been given intelligible, easily accessible, and concise information about the risks and benefits of giving consent. This raises the issue highlighted in *Clarkson v OpenAI*, where non-consenting adults and even children were featured in deepfakes without their consent, showcasing the prevalence of unauthorised use in this context.²⁴

If there is no consent to use personal data for deepfake creation or dissemination, then a careful assessment arises of the potential risks to individual rights and freedoms through the use of deepfakes. Consequently, assessing whether deepfake developers, creators, and distributors can rely on 'legitimate interest' as a legal basis under GDPR's Article 6 (1)(f) requires carefully weighing the potential benefits of processing personal data for deepfakes against potential harms to individual rights and freedoms. While the article itself does not explicitly list relevant factors, the CJEU's *Rīgas*²⁵ ruling emphasises the need to strike a balance between personal data processing and protecting users' rights and freedoms.

Arguably, deepfakes used for artistic, satirical, or fictional purposes (Recital 134 AI Act) could fall under creators' freedom of expression and art/science (Articles 11 and 13 of the Charter). However, creating electoral misinformation, extortion material, or AI-generated sexual abusive content would clearly require prioritising the protection of victims' privacy and data protection rights (Articles 7 and 8).

Furthermore, following the precedent set in the *Rīgas*²⁶ case, the justification for using deepfakes on data protection grounds might hinge on a careful analysis of their necessity, weighing their potential benefits against the risks they pose. On the one hand, deepfake detection, despite limitations, can shield individuals from harm. It identifies manipulated content, preventing the spread of damaging misinformation and protecting individuals' reputations and privacy. It also safeguards against data theft by detecting fabricated identities used in cyberattacks. Furthermore, in sectors like finance and healthcare, where data accuracy is crucial, it ensures decisions are based on genuine information. However, several challenges require attention. First, deepfake models may discriminate against certain individuals due to biased training data. Second, regarding transparency, lack of explanation in deepfake detection methods reduces trust and hinders legal applications. Third, in terms of accuracy, these tools struggle to detect all manipulated content and differentiate them from genuine variations (EDPS 2023).

4.2. Limited exceptions, big challenges: balancing deepfake and sensitive data with article 9 GDPR

Adding complexity to the landscape, the CJEU ruling in *Meta v Bundeskartellamt* expands GDPR protections, potentially subjecting deepfakes to stricter rules. Personal information that might reveal sensitive data (like political views, sexual orientation) is now potentially subject to Article 9 protections, even if it does not directly mention them.²⁷ This means processing data to create deepfakes could trigger stricter privacy requirements, even if the data seems harmless. Developers must then navigate the limited exceptions within

Article 9(2). Unlike Article 6(1)(f) GDPR, Article 9(2)'s most readily available option for exemption would be explicit user consent, as it lacks a balancing test like the one employed in *Rigas*²⁸ between societal benefits and individual rights. This makes justifying deepfakes, even those with potential positive intentions, extremely challenging. The research exemption does not apply to commercial deepfakes, further limiting options. The GDPR's ambiguous data privacy framework necessitates a new, well-protected exemption in Article 9, similar to the AI Act's Article 10(5). Such an exemption, aiming to balance the benefits of responsible AI development for media, advertising, and entertainment with individual privacy, could address the challenges posed by deepfakes under the GDPR (Novelli et al. 2024, 11).

4.3. Data subject protections against deepfakes in the GDPR

Beyond legal justifications, the GDPR offers additional protection for deepfake victims. Article 22 grants individuals the right to object to automated decisions significantly impacting their lives, based solely on their likeness in malicious deepfakes. This includes blackmail, financial scams, abuse depictions, and election misinformation. Consider a deepfake used for blackmail. Article 22(2) GDPR shields the victim from automated AI decisions (e.g. credit denial) solely based on the manipulated video, without their consent (highly unlikely in such cases). However, existing alternatives like contracts (impractical) or EU/Member State laws lack stringent safeguards. While the EU AI Act's Article 50 requiring detection and disclosure by deepfake providers and deployers satisfies the 'authorised by EU law' requirement, it falls short of establishing truly effective safeguards for fundamental rights and freedoms.

The AIA needs to be strengthened to adequately protect individuals from harmful automated decisions based on deepfakes. The CJEU's *SCHUFA (Scoring)* ruling emphasised this. The CJEU stressed that examining individuals' right not to be subject to automated decision-making and profiling under Article 22(1) GDPR; raised concerns about discrimination in automated systems analysing sensitive data like political views or sexual orientation (Recital 71, GDPR).²⁹ This emphasises the need for robust safeguards, like transparency, minimising bias in algorithms, and secure data handling. Additionally, individuals must have the right to challenge decisions, express their views, and request human intervention, as outlined in Recital 71 GDPR.³⁰

Furthermore, while the GDPR grants individuals rights over their personal data, including the right to be forgotten (Article 17) and the right to rectification (Article 16), enforcing these rights poses significant challenges in the context of deepfakes. Deepfakes, AI-generated videos or audio manipulated to resemble someone else, present unique complexities. Unlike traditional data, personal information in deepfakes is intricately woven into the fabricated content, making it exceedingly difficult to isolate and remove specific data points, especially for unstructured data like voice and facial features (Novelli et al. 2024, 14). Additionally, the viral nature of deepfakes, easily disseminated and replicated across platforms and jurisdictions, renders removal efforts challenging and potentially ineffective (Brown et al. 2022).

The case of *Clarkson v OpenAI* further exemplifies these difficulties, highlighting the limitations of current frameworks in applying the 'right to be forgotten' to generative AI models like ChatGPT. The privacy policy in question failed to address how data, once

integrated, could be truly removed, underscoring the need for more robust legal and technical solutions to tackle the challenges posed by deepfakes.³¹

4.4. Outsmarting deepfakes: the advantage of synthetic data

The EU's AI Act strikes a balance between fair and unbiased AI and personal data protection. While it allows 'high-risk' AI developers to access sensitive data under strict conditions for bias detection (only when absolutely necessary and after exploring alternatives like 'synthetic data' – Article 10(5)(a), it currently does not address deepfakes. This is concerning because deepfakes, despite not being classified as high-risk, can distort reality, spread misinformation, and fuel discrimination if not developed responsibly.

To address this challenge, organisations like the Confederation of European Data Protection Organisations propose promising solutions using privacy-enhancing technologies like synthetic data. This approach involves generating realistic simulated data instead of relying on real people's information, offering a potential route to mitigate the risks associated with deepfakes. First, by avoiding sensitive real-world data, synthetic data helps prevent AI models from inheriting and amplifying societal biases, thereby minimising the risk of discriminatory deepfakes. Second, it strengthens privacy and security by reducing the dependence on personal information, lowering the risk of privacy breaches and unauthorised use. Third, the generation processes for synthetic data can be more transparent and easier to explain, addressing concerns about the 'black box' nature of some AI systems (CEDPO 2023, 21). It is crucial to emphasise that this approach also eliminates the 'overfitting' problem, where a system excels on training data but fails to generalise to new content (Syntheticus 2023). While not a silver bullet, synthetic data represents a significant step towards developing responsible deepfakes that comply with the AI Act and GDPR's requirements and respect privacy and security.

5. Finding balance: a proposed framework for responsible deepfake use

Article 7(1) of the AIA empowers the Commission to add or change areas or applications of high-risk AI systems to Annex III through delegated legislation under Article 97. This can occur under two conditions. Firstly, if the AI system is designed for uses listed in the specific categories of Annex III, such as influencing elections or voting behaviour. Secondly, if it poses a risk to fundamental rights and safety equal to or exceeding those of existing high-risk systems in Annex III. Furthermore, Article 112(10) mandates the Commission to propose amendments to the AIA as needed, considering technological advancements, the impact of AI systems, and the evolving information society.

Deepfakes, once science fiction, are now eroding trust in elections, enabling extortion, and putting children at risk through AI-generated abuse material. Considering these alarming threats, this paper proposes crucial amendments to the EU AI Act to safeguard fundamental rights and ensure a safer digital future.

Firstly, the AIA must mandate the use of 'structured synthetic data' for deepfake detection. This specialised data format empowers AI systems to better identify and flag manipulative content, protecting fundamental freedoms and enhancing overall security. Secondly, the paper proposes classifying AI intended for deepfake electoral disinformation, extortion, and AI sexual abuse content as 'high-risk' AI. This categorisation would

trigger stricter regulatory oversight and controls, reflecting the significant potential for harm and violation of rights these practices pose.

In *Poland v Parliament and Council*, AG Saugmandsgaard Øe explained that uploading various forms of content constituted an exercise of the right to freedom of expression and information, protected under Article 11 of the Charter. Additionally, the AG recognised that for content involving artistic expression uploaded by users, it also fell under the protection of freedom of the arts, enshrined in Article 13 Charter and Article 10 Convention.³² Saugmandsgaard Øe observed that the right enshrined in Article 11 Charter corresponded to that contained in Article 10 of the Convention. He emphasised that, under Article 52(3) of the Charter, these two rights were identical in meaning and scope. Accordingly, the AG stressed that interpreting Article 11 Charter requires consideration not only of Article 10 of the Convention but also of the relevant case-law of the Strasbourg Court.³³

In *Poland v Parliament and Council*, the CJEU, following the AG's recommendations, emphasised that the EU Charter guaranteed freedom of expression, including holding opinions, receiving information, and sharing ideas freely, without government interference or borders.³⁴ As confirmed by the Charter's explanations and Article 52(3), Article 11 (freedom to receive and impart information) mirrored Article 10 ECHR (freedom of expression), in both definition and scope. The CJEU noted that sharing information online, through platforms like social media, enjoyed the protection of free speech, enshrined in Article 10 of the Convention and Article 11 of the Charter.³⁵

The CJEU elaborated that, according to ECtHR's case law, Article 10 ECHR protects freedom of expression and information, encompassing both the content and means of dissemination. Hence, any restriction on dissemination hinders the right to receive and impart information.³⁶ Furthermore, referencing concerns raised in *Yildirim v Turkey*³⁷ by the ECtHR, the CJEU emphasised the need for a particularly rigorous legal framework for prior restraints on online content dissemination due to the significant risk they pose to freedom of expression and information.³⁸

In *Poland v Parliament and Council*, the AG Saugmandsgaard Øe concluded that pursuant to Article 10(2) ECHR and the ECtHR's case-law, a restriction on freedom of expression was only allowed if it, firstly, was 'prescribed by law', secondly, pursued one or more legitimate aims outlined in paragraph 2 and, lastly, was 'necessary in a democratic society'.³⁹

6. Assessment of applicability and compliance with articles 8 and 10 of the ECHR

6.1. Unpacking the 'in accordance with law' requirement: defining the boundaries of permissible deepfakes

The ECtHR has ruled that for any interference with the right to privacy and freedom of expression under Articles 8 and 10 of the Convention to be considered 'in accordance with the law', it must satisfy three criteria: firstly, it must be based in domestic legislation; secondly, this law must also be accessible; and lastly, it must additionally satisfy the Strasbourg Court's principles of foreseeability and rule of law.⁴⁰ The requirement that the interference be based in domestic legislation is not difficult to meet as the AIA, which is statutory law, and the ECtHR facial recognition ruling in *Glukhin v Russia*,⁴¹ offer this.

However, concerning the second and third criteria, this section argues that the EU AI Act's deepfake provisions could be inconsistent with the Strasbourg Court's principles of accessibility, foreseeability, and the rule of law. This potential inconsistency could violate the first criterion of the Court's three-part, non-cumulative test under the right to privacy (Article 8(2)) and freedom of expression (Article 10(2)) ECHR.

6.1.1. *The accessibility principle*

In terms of the accessibility principle, Strasbourg Court case-law has established that legislation must be adequately accessible, allowing individuals to clearly understand the legal norms applicable to their situation.⁴² However, unsettling, the EU AI Act and the EU DSA framework present conflicting classifications for AI used in electoral disinformation. Recital 132 of the AIA identifies concerns around 'certain' AI systems interacting with individuals or generating content, highlighting their 'specific risks' like deception (e.g. deepfakes). However, it also implies that deepfake disinformation related to elections falls under the 'systemic risk' category within the DSA framework (Recitals 120, 136). Confusingly, Recital 62 contradicts this by classifying AI intended to influence elections as 'high-risk.' This glaring inconsistency raises crucial questions: How should AI used for electoral disinformation be categorised? Should it be considered a 'systemic risk' or 'high-risk'? Moreover, the distinction is vital as high-risk AI faces stricter regulations under the EU AI Act. What is clear, however, is that to respect the right to freedom of expression under Article 10 ECHR, ECtHR case-law⁴³ has observed that law tackling electoral disinformation should normally only target knowingly false information intended to manipulate voters or erode the rights of others. Conversely, legislation targeting misinformation should be discouraged, especially if it incorporates criminal sanctions (Shattock 2022, 25).

Additionally, two further questions arise. Firstly, how do the regulations apply to AI systems specifically used for malicious deepfakes related to extortion and sexual abuse? Secondly, should such AI systems be considered a 'specific risk,' 'systemic risk,' or 'high-risk'? These inconsistencies and ambiguities demand clear and consistent regulations; they even contradict ECtHR⁴⁴ and CJEU⁴⁵ caselaw, which is particularly troubling.

To ensure clarity and protect human rights, the AIA should clearly define 'certain' AI systems prone to deception and impersonation, like deepfakes. It should also align risk classifications and mandate transparency for all high-risk AI, not just electoral disinformation. Thus, to comply with the accessibility principle, the AIA could require conformity assessments (Article 6(4)) and public EU database registration (Art. 49(2), Art. 71), not only for deepfake electoral disinformation systems but also for AI used for extortion and sexual abuse. As the ECtHR succinctly put in *Węgrzynowski and Smolczewski v Poland*, serving billions of users globally, the web was not and will never be subject to the same control and rules. The risks of harmful content must undoubtedly be revised according to the technology's characteristics to ensure the enjoyment of human rights and freedoms.⁴⁶ The AIA's inconsistent classification of electoral disinformation leaves providers and deployers struggling to understand how other potentially high-risk AI, like systems used for extortion or sexual abuse, will affect them. Thus, this arguably violates the ECtHR accessibility principle enshrined in Articles 8(2) and 10(2) ECHR.

6.1.2. The foreseeability principle

As far as the foreseeability principle is concerned, the ECtHR has established that a rule cannot be considered law unless it is clear enough for individuals to reasonably understand the consequences of their actions.⁴⁷ Article 50 AIA mandates some transparency measures for deepfakes (e.g. disclosure, tags). However, it fails to address the crucial role communication platforms could play in labelling and monitoring harmful content, raising concerns about the Act's compliance with established platform accountability principles outlined in ECtHR⁴⁸ and CJEU⁴⁹ caselaw. To ensure deployers understand the consequences, the AIA should clarify that transparency does not translate into general monitoring requirements for platforms, potentially violating freedom of expression as evidenced by CJEU *Poland v Parliament and Council*⁵⁰ and *Glawischnig-Piesczek*.⁵¹ While preventive measures might be appropriate, platforms should not be burdened with individual content assessments, especially considering copyright complexities.

Adding to the concerns, the AIA introduces exemptions and limited transparency requirements for specific types of deepfakes. Law enforcement deepfake detection models, initially classified as high-risk, are now entirely exempt from transparency requirements, raising concerns about potential misuse and infringement on individual rights, even for law enforcement purposes (Labuz 2023, 14). Similarly, the limited transparency imposed on 'creative' deepfakes (Article 50(4)) weakens user awareness and hinders the identification of harmful content disguised as creative work (Ajder and Glick 2021). While arguments exist for these exemptions (e.g. protecting law enforcement methods, artistic freedom), the potential risks to individual rights and effective public oversight should not be ignored.

Additionally, troubling gaps exist in enforcing transparency obligations for deepfakes under the EU AI Act. While individuals can report non-compliance with Article 50(4) reporting requirements (Article 85a), Article 99 curiously lacks any specific penalties for deployers who fail to comply. This lack of defined consequences creates uncertainty for deployers and undermines the effectiveness of the transparency measures. Contrasting with established ECtHR⁵² principles, notably Article 10 ECHR on freedom of expression, which allows for 'penalties as are prescribed by law,' the AIA lacks such provisions for deployer non-compliance with Article 50(4)'s transparency obligations (EP 2021, 38). This absence of foreseeable consequences can be interpreted as violating the ECtHR's foreseeability principle under Articles 8(2) and 10(2).

6.1.3. The rule of law principle

In upholding the rule of law principle, the ECtHR has emphasised the need for initial state oversight and safeguards for intrusive measures under Articles 8(2) and 10(2) ECHR.⁵³ Recognising this, Article 50(7) of the AI Act empowers the AI Office to enforce regulations against deepfakes through a two-pronged approach, starting with voluntary codes and escalating to Commission-stricter rules if needed. However, the AIA's disregard for legal precedents set in *Strasbourg*⁵⁴ and *Luxembourg*,⁵⁵ particularly by failing to mandate prior authorisation for invasive tracking tools like fingerprinting and metadata, is deeply worrying.

As demonstrated in *Clarkson v OpenAI*, this allows companies to build detailed profiles of users' online activity (e.g. browsing history, device information) without knowledge or

consent, potentially enabling targeted advertising, discrimination, and personalised scams.⁵⁶ This directly contravenes the right to privacy enshrined in Article 8 ECHR. Furthermore, the lack of transparency regarding data scraping methods rendered user consent meaningless, effectively subjecting them to secret monitoring, and profiling.⁵⁷ This lack of control over personal data and potential manipulation of user behaviour raises concerns about its impact on democratic processes and freedom of expression, potentially violating Article 10 ECHR.

The European Data Protection Supervisor (EDPS 2010, 8), the AG opinion in *Promusicae*⁵⁸ and the CJEU in *Tele2/Watson*,⁵⁹ all stressed the importance of prior review by independent authorities like the AI Office to protect individuals' privacy and data protection rights. The AIA's lack of such safeguards, particularly for invasive tracking tools, creates a cause for alarm regarding its effectiveness in preventing discriminatory profiling, algorithmic bias, and other potential privacy harms.

Moreover, the 'right to be forgotten' principle, enshrined in the GDPR (Article 17) and supported by CJEU⁶⁰ and ECtHR⁶¹ safeguards, grants individuals control over their personal data, including the right to request its deletion. This is crucial in the context of deepfakes, where data integration raises concerns about potential misuse. As seen in *Clarkson v OpenAI*, the lack of deletion assurances for integrated data violates the 'right to be forgotten,' preventing individuals from exercising control over their data.⁶² This potential for misuse is further amplified by the AIA's lack of safeguards for minors, as highlighted in OpenAI's failure to ensure parental consent or erasure of their information.⁶³

Significantly, similar to the unclear erasure procedures criticised in *Kurić v Slovenia*, the EU AI Act's lack of transparency regarding data processing and control could violate individuals' right to control their personal data under Article 8 ECHR.⁶⁴ This, coupled with the absence of prior authorisation for invasive tracking tools, creates a situation where individuals have limited control over their data and its potential use in deepfakes. It thus directly contradicts the ECtHR's rule of law principle enshrined in Articles 8(2) and 10(2).

6.2. Legal justifications for deepfakes

Pursuant to Article 8(2) and Article 10(2), state authorities can refer to several clearly defined legitimate grounds to justify the limitation of the right to privacy and freedom of expression under the Convention. These include grounds such as securing the state's security, the protection of citizens, the economic well-being of the country, the deterrent of crime or disorder, and the safeguarding of the rights and freedoms of others.⁶⁵ The second requirement of the Court of Strasbourg's non-cumulative test, which requires that the interference achieves a legitimate aim, is typically not a difficult hurdle for states to overcome. It is possible that systems that use AI to detect deepfakes could be deployed to prevent crime or disorder and protect the reputation and rights and freedoms of others, as exemplified by the ECtHR's findings in the ruling of *Glukhin v Russia*, which involved the Russian government's use of facial recognition.⁶⁶

6.3. Striking a balance: 'necessity' and 'proportionality' in deepfake regulation

The next issue to be examined in this paper is to what extent the provisions governing the use of deepfakes in the EU AI Act would meet the third requirement of the Strasbourg

Court's three-part, non-cumulative test. The ECtHR has ruled that measures are considered justified in a democratic society if they address a 'pressing social need' and are proportionate to the legitimate aim pursued.⁶⁷ In addition, the Strasbourg Court has noted that the reasons provided by the state to justify the measures must be 'relevant and sufficient'.⁶⁸ Yet, even though state authorities have some flexibility margin, the final determination of whether such measures remain necessary and proportionate is subject to the scrutiny of the ECtHR in Strasbourg.⁶⁹ This section will argue that the AIA's deepfake provisions are unjustified due to their excessive restrictions on privacy and expression, resulting in violations of the Strasbourg Court's necessity and proportionality principles.

6.3.1. The necessity principle

Regarding the first principle, Strasbourg caselaw has firmly established that the level of disruption caused by a measure is crucial when evaluating necessity under Articles 8(2) and 10(2) ECHR.⁷⁰ Article 10(5)(a) AIA mandates exploring alternative methods like structured synthetic data before processing sensitive data, for bias detection in AI high-risk systems. This aligns with the GDPR's principles of data protection by design and by default and data minimisation (Recital 69).

A concerning aspect of the EU AI Act is the exemption for certain deepfake creators, like those using them for electoral disinformation, from conducting mandatory fundamental rights impact assessments (Article 27, Recital 96). These assessments evaluate potential risks to fundamental rights like privacy and freedom of expression. This exemption is particularly problematic for high-risk categories, as it allows creators to potentially manipulate voters without formally considering the potential impact on these rights. While they are still required to assess data processing necessity and proportionality through data protection impact assessments (Article 26(9), Article 27(4)), even more concerning is the complete exemption for creators of deepfakes used for extortion and AI-generated child abuse content from both types of assessments. This significant loophole creates a breeding ground for highly harmful content, leaving individuals vulnerable to human rights violations.

Furthermore, the AIA's lack of clear limitations on tracking tools also creates a regulatory gap, enabling intrusive data collection. Research suggests alternative approaches, like prohibiting watermark services from storing or selling data (Brookings 2024), aligning with Strasbourg⁷¹ and Luxembourg⁷² precedents advocating for less invasive measures. Additionally, the Global Privacy Control empowers users to directly manage data sharing preferences, including cookies, sales, and targeted advertising (Usercentrics 2023).

To effectively protect individuals and uphold GDPR rights in the absence of adequate regulations, deepfake detection system providers targeting harmful content should prioritise data minimisation techniques like structured synthetic data at the AI design stage. This not only addresses data minimisation concerns and mitigates bias, but also effectively handles limited data availability, making it a crucial approach for responsible deepfake detection (Intel. 2023). However, even a small false positive rate can create significant harm, as highlighted by the CJEU AG opinion in *Poland v Parliament and Council*. Thus, if using structured synthetic data leads to a significant increase in false positives, such data filtering should be prohibited.⁷³ Intel's real-time FakeCatcher demonstrates this feasibility,

achieving 96% accuracy within milliseconds without privacy-invasive measures (Intel. 2022). Therefore, the AIA demonstrably fails to satisfy the ECtHR necessity principle under Articles 8(2) and 10(2), prioritising provider interests over minimising data collection and protecting fundamental rights.

6.3.2. The proportionality principle

In applying the proportionality principle, the ECtHR has consistently emphasised that upholding individual protection requires robust safeguards against arbitrary interference.⁷⁴ However, Recital 133's requirement for deepfake providers to implement invasive tracking tools like watermarks, metadata tags, and fingerprints significantly undermines this principle. This flexible approach fundamentally contradicts established legal principles demanding relevant, minimal, and transparent data collection, as enshrined in Article 5 of Convention 108, and exemplified by ECtHR⁷⁵ and CJEU⁷⁶ rulings.

Notably, website tracking tools collect not just basic information, but also user behaviour data revealing sensitive details like political views and sexual orientation (CJEU *Vryiausioji*⁷⁷ and *Meta v Bundeskartellamt*).⁷⁸ This inferred data, derived from online behaviour, perpetuates discriminatory practices like unequal job opportunities or targeted political ads (illustrated by the Cambridge Analytica scandal),⁷⁹ furthermore amplifying the spread of misinformation and blackmail. These practices silence dissent and erode trust in information.

Terrifyingly, a well-intentioned rule intended to curb deepfakes (Recital 133) could backfire spectacularly. Empowered with mass surveillance capabilities, deepfake providers could jeopardise personal information and privacy, fuelling discrimination through manipulated media like videos discrediting minority candidates. Making matters worse, the AIA not only ignores stringent data quality standards for high-risk AI like deepfakes (Recital 67), but also overlooks critical scenarios like electoral disinformation campaigns, targeted extortion tactics, and fabricated child abuse content.

To ensure individual protection and align with established legal principles, the AIA should require deepfake providers to prioritise design-stage trade-offs. This includes balancing statistically accuracy with data minimisation, preventing discrimination, and striking a fair balance between explainability, commercial secrecy, and security. While 'statistical accuracy' concerns the AI's inherent performance, deepfake detection systems need not be 100% accurate to comply with the accuracy principle (ICO 2023, 24, 39). However, echoing the concerns raised by the CJEU AG in *SCHUFA (Scoring)*, deepfake detection providers should be required to provide meaningful information about the system's logic. This includes comprehensive explanations of the detection techniques used and the rationale behind specific outcomes,⁸⁰ especially when tackling sensitive issues like electoral disinformation, AI sexual abuse material, and extortion. In these contexts, knowing the reasons for significant or insignificant 'false positive' rates are crucial. As the AG emphasised in *SCHUFA (Scoring)*, individuals should also be given general information about the decision-making factors and their weight, empowering them to contest decisions and exercise their GDPR right to avoid automated processing solely based on profiling.⁸¹

Furthermore, *Clarkson v OpenAI* demonstrates how generative AI models can be exploited for algorithmic discrimination and predatory advertising, perpetuating harm against vulnerable populations like children, refugees, or minority groups, even under

claims of 'absolute secrecy' regarding data practices.⁸² Additionally, OpenAI's development raised concerns about potential autonomous weapons, posing a grave threat to domestic and international security.⁸³ Unfortunately, the EU AI Act does not address these crucial issues. It neither requires deepfake providers to implement statistically accurate and non-discriminatory systems nor mandates a fair balance between explainability, accuracy, security, and commercial secrecy. These shortcomings weaken the protection of individual rights and arguably violate the ECtHR's proportionality principle under Articles 8(2) and 10(2).

Regarding the proportionality principle, the ECtHR has stressed that monitoring and technical measures must balance competing interests fairly.⁸⁴ However, the AIA, in Recital 107, mandates general-purpose AI providers to publicly share high-level overviews of their training data, while seemingly contradicting established legal precedents like ECtHR⁸⁵ and CJEU⁸⁶ by only protecting copyright, trade secrets, and confidential information. This raises concerns about potential violations of crucial protections for user privacy (Article 7), data protection (Article 8), and intellectual property rights (Article 17(2) Charter). Furthermore, the evidently greater weight given to intellectual property rights over AI companies' freedom to conduct business (Article 17(2) vs Article 16 Charter) further highlights potential proportionality and fundamental rights balancing issues.

Firstly, OpenAI's technical report reveals how ChatGPT-4 was trained on publicly available information like online data and third-party provider licenced data (OpenAI 2023, 2). Importantly, the CJEU case *Poland v Parliament and Council* emphasised the need for balancing interests under Article 16 of the EU Charter, allowing providers flexibility in technical measures based on their capabilities.⁸⁷ However, as highlighted in *Clarkson v OpenAI*, the source of OpenAI's data raises concerns.⁸⁸ This case revealed the commercial misappropriation of the 'Common Crawl' database, involving the extraction of millions of users' personal information (like emails, social media posts, browsing history) without their knowledge or consent, potentially violating both privacy and property rights and constituting theft.⁸⁹

Furthermore, according to legal documents regarding the *Clarkson v OpenAI* case, OpenAI used training datasets for ChatGPT that included not only user-prompted information (e.g. questions, creative prompts) but also information like contact details, account information, IP addresses, login credentials, cookies, and analytics.⁹⁰ This raises significant concerns about data privacy, as highlighted in the case. Additionally, the case warned that the vast amount of individualised and sensitive data, encompassing audio, visual, and personal preference data, could fuel the widespread creation of harmful 'deepfakes,' such as impersonation videos or fake news articles.⁹¹

Moreover, the AI Act exempts free and open-source AI systems from transparency requirements imposed on general-purpose AI models under Recital 104 for reasons of proportionality and business freedom. While this aims to incentivise open development, it creates a concerning regulatory gap. Platforms like CivitAI, known for its 'bounty' feature allowing users to request deepfakes of specific individuals, can operate outside the AIA's regulations unless deemed systemic risk. This loophole raises critical questions about the potential for unregulated companies to exploit it for harmful purposes like creating non-consensual deepfakes, identity theft, or disinformation campaigns. Additionally, it allows them to gain unfair advantages over legitimate businesses operating responsibly within the AIA's regulations, potentially distorting competition. Ultimately, as the AIA's current

form fails to strike a fair balance between users' privacy, data protection, intellectual property rights, and AI companies' business freedom, it risks violating the ECtHR proportionality principle under Articles 8(2) and 10(2).

7. Deeper dive: scrutinising the key findings

7.1. *Strasbourg and Luxembourg echoes, aligning findings with ECHR and EU Charter legal frameworks*

The findings of this research align with the rulings of the Strasbourg and Luxembourg courts. In *Glukhin v Russia*, the ECtHR emphasised that national laws on data gathering and processing must have specific and transparent rules specifying the application and scope of measures, along with minimal safeguards.⁹² The ECtHR strongly doubted whether the domestic framework governing the use of AI-based facial recognition satisfied the 'quality of law' condition.⁹³ However, it highlighted that its role was strictly to determine whether the applicant's data processing was 'necessary'.⁹⁴ It found that the deployment of facial recognition to detect *Glukhin* violated Article 8 ECHR.⁹⁵

Similarly, in *Poland v Parliament and Council*, observing *Promusicae*,⁹⁶ the CJEU confirmed that while specifically targeted filtering and blocking measures that adequately distinguish between lawful and unlawful content might be deployed, general monitoring obligations imposed on providers to take preventive measures against future infringements were indeed prohibited.⁹⁷ Moreover, when implementing such measures, domestic legislation needed to allow a fair balance to be struck between all relevant interests, while also ensuring respect for Charter rights, including the principle of proportionality.⁹⁸

Notably, the paper's findings could also have broader implications for society, the environment, and the economy.

7.2. *Decoding sustainability, Google's AI guides eco-friendly policy frameworks*

Regarding environmental issues, research discloses that training ChatGPT-3 needed 1,287 megawatt hours of electricity, which equates to the annual electricity consumption of 121 US homes. It also generated 552 tons of carbon dioxide equivalent (CO₂e), which compares to the emissions from driving a car 1.3 million miles (Patterson et al. 2021, 7). However, problematically, neither OpenAI's technical report (OpenAI 2023), nor *Clarkson v OpenAI*⁹⁹ unmasked the carbon footprint of developing ChatGPT-4. While one large-scale generative AI model would not cause much environmental damage, if numerous companies developed models with slight differences for various purposes – such as detecting extortion, electoral disinformation, and AI sexual abuse content – each used by millions of users, the high demand for energy could undeniably become unsustainable (Saenko 2023).

Article 95(2)(b) of the AIA requires the AI Office and Member States to facilitate the creation of voluntary sustainability codes of conduct. These codes aim to assess and minimise the environmental impact of AI systems throughout their lifecycle, through energy-efficient techniques in programming, design, training, and use. However, evaluating a system's CO₂e emissions through traditional environmental impact assessments poses challenges. Quantifying all necessary information – particularly regarding hardware,

datacenters, energy mix, and subsequent data disclosure – can be difficult due to limited availability or accessibility (Patterson et al. 2021).

To actively encourage the development of voluntary codes that effectively minimise AI's environmental impact through energy efficiency, the AI Office and Member States may consider supporting and adopting Google's leading approach. Google's business model has consistently prioritised improving the energy efficiency of algorithms, software, hardware, and datacenters. For example, Google Cloud allows clients to choose datacenters based on their CO₂e footprint and publishes regular updates on the level of carbon-free energy and gross CO₂e emissions for each facility (Google Cloud 2023). By embracing Google's approach, the AI Office and Member States can leverage established best practices and encourage wider adoption of energy-efficient AI development across the EU.

7.3. Unveiling and countering deepfakes, a multi-pronged approach

While the EU AI Act's focus on technical solutions (Article 50(2)) deserves recognition, tackling deepfakes demands a broader approach. Research on integrated detection methods offers promise, as seen in (Thi Nguyena et al. 2022, 13). However, relying solely on notice-and-takedown/staydown systems raises censorship concerns and verification challenges (Romero-Moreno 2019; 2020).

Understanding the 'why' behind deepfakes is crucial, not just for detection. Analysing motivations and public reactions (Intelligencer 2019) can inform targeted interventions and educational strategies to help users critically evaluate online content. However, human limitations in detection (The Guardian News, August 2, 2023) highlight the need for more.

Combating deepfakes requires a multi-pronged approach, with user empowerment playing a vital role. Firstly, platforms can empower users through: interactive tutorials identifying deepfake cues; engaging games discerning real from fake content; fact-checking resources enabling easy access to verification tools; and media literacy education integrating critical thinking skills (MIT Center for Advanced Virtuality 2021).

Secondly, content verification is essential. Leveraging technologies such as blockchain (Hasan and Salah 2019), or privacy-enhancing digital fingerprints can transparently trace content origins (Barrington et al. 2023).

Thirdly, collaboration between stakeholders is vital. Successful projects require a multi-disciplinary approach, including AI developers, consumer groups, civil society, organisations representing victims, executives from small, medium, and large businesses, as well as scientists and researchers (AIA's Recitals 142, 165). Partnerships can accelerate progress through shared knowledge and data (Recital 74).

Furthermore, prioritising ethics is paramount. Open dialogue and clear frameworks are essential as deepfake technology evolves, focusing on issues like transparency, accountability, potential harms, and responsible AI development.

Lastly, ongoing research and development in deepfakes detection, prevention, and user education are critical.

8. Concluding remarks: key takeaways on deepfakes and the EU AI Act

This paper has examined the compatibility of the EU AI Act's deepfake provisions with the right to privacy and freedom of expression (Articles 8 and 10 of the Convention and the GDPR) for

both AI providers and users. It has argued that implementing these provisions without safeguards could violate these rights. Consequently, the paper has proposed procedural safeguards necessary to ensure compliance with Articles 8 and 10 ECHR and the GDPR. Only through incorporating these safeguards, potentially via delegated acts under Article 7(1), can the AIA's deepfake provisions effectively regulate without infringing on fundamental rights.

Firstly, the EU's 'high-risk' label fuels confusion. It fails to clarify how electoral rules apply to harmful AI like extortion or sexual abuse bots. This ambiguity stifles both prediction and compliance, threatening accessibility. The AIA needs a clear definition of high-risk AI, especially deepfakes, to ensure clear understanding and consistent application.

Secondly, vague platform roles for deepfakes (disinformation, extortion, sexual abuse content) in the EU AI Act throttle response and violate foreseeability. Unclear consequences for transparency breaches (Article 50(4)) further hinder enforcement and accountability, contradicting EU's values. The AIA must clearly define platform responsibilities for detecting and flagging harmful deepfakes, with appropriate sanctions for violations.

Thirdly, the EU AI Act's lax approach to tracking and deepfakes undermines the rule of law. The fix: demand prior AI Office approval and individual empowerment for deepfakes in disinformation, extortion, and child abuse, while empowering individuals with notification, deletion, and parental consent rights. These safeguards ensure accountability and protect individuals from deepfake threats.

Furthermore, the EU AI Act prioritises convenience over safety, by exempting some content creators from assessments, while grilling others to scrutiny. It even lacks privacy-friendly tracking, putting users at risk. To fix this, mandate necessary assessments for all deepfakes, especially high-risk ones, and use synthetic data instead of real user information in sensitive areas.

Additionally, the EU AI Act's deepfake detection fails to meet fair detection Article 5's Convention 108 standards. Biased and opaque algorithms lack accuracy and accountability. To uphold proportionality, the AIA should mandate fair detection principles and require explainable algorithms balancing security and business needs.

Lastly, the AIA forces AI providers to share broad data, prioritising commercial interests over user privacy, data protection, intellectual property rights, and even impacting business freedom. This imbalance demands adjustments to ensure proportionality. The AIA should recognise the interests of all stakeholders, not just rightsholders, and refine data sharing for transparency.

Urgent action is vital. The EU AI Act's deepfake provisions, despite careful crafting, lack crucial safeguards, posing a significant risk to fundamental rights protected by Articles 8 and 10, ECHR and GDPR. Implementing them without safeguards could stifle responsible AI development, potentially leading to widespread misinformation and manipulation, and harm individual rights. Policymakers must mandate clear definitions, transparent and accountable oversight, and robust safeguards for both users and providers. Only then can the AIA effectively regulate deepfakes without becoming a weapon against the very freedoms it seeks to safeguard.

Notes

1. *PM et al v OpenAI LP*, 3:23-cv-03199 (US District Court, N.D. Cal. 2023) [219].
2. *Ibid.* [220].

3. *Ibid.* [222], [223].
4. *Ibid.* [223].
5. *Ibid.* [224].
6. *Ibid.* [225].
7. AG opinion in *C-401/19 Poland v Parliament and Council* [2021] ECLI:EU:C:2021:613 [214].
8. *C-314/12 UPC Telekabel Wien GmbH v Constantin FilmVerleih GmbH and Wega Filmproduktionsgesellschaft GmbH* [2013] EU:C:2014:192 [52].
9. *PM et al v OpenAI LP* (N.D. Cal. 2023) [226].
10. *Halet v Luxembourg* App no 21884/18 (ECtHR, 14 February 2023).
11. *PM et al v OpenAI LP* (N.D. Cal. 2023) [222] [223].
12. *Ibid.* [222], [223], [224], [225].
13. *Ibid.* [220].
14. *Ibid.* [222]-[225].
15. *Ibid.* [222], [223].
16. See e.g. Case T-250/15 *Speciality Drinks Ltd v European Union Intellectual Property Office – William Grant (CLAN)* ECLI:EU:T:2016:678 [26].
17. *C-324/09 L'Oréal SA and others v eBay International AG and others* [2011] ECR I-0000 [122].
18. *Joined Cases C-682/18 and C-683/18 Frank Peterson v Google LLC and Others and Elsevier Inc. v Cyando AG* ECLI:EU:C:2021:503 [116].
19. *C-210/16 Unabhängiges Landeszentrum für Datenschutz Schleswig-Holstein v Wirtschaftsakademie Schleswig-Holstein GmbH* [2018] EUECJ [36]-[39].
20. *C-698/15 Tele2 Sverige AB v Postoch telestyrelsen* [2016] All ER (D) 107 (Dec) and *Secretary of State for the Home Department v Tom Watson* [2016] All ER (D) 107 (Dec) [123].
21. *C-634/21 SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:957 [56], [59], [66].
22. *C-673/17 Bundesverband der Verbraucherzentralen und Verbraucherverbände – Verbraucherzentrale Bundesverband e.V. v Planet49 GmbH* [2019] EU:C:2019:246 [61].
23. *C-61/19 Orange Romania SA v Autoritatea Națională de Supraveghere a Prelucrării Datelor cu Caracter Personal (ANSPDCP)* [2020]ECLI:EU:C:2020:901 [8].
24. *PM et al v OpenAI LP* (N.D. Cal. 2023) [222], [223].
25. *C-13/16 Valsts policijas Rīgas reģiona pārvaldes Kārtības policijas pārvalde v Rīgas pašvaldības SIA "Rīgas satiksme* [2017] 4 WLR 97 [28]-[32].
26. *Ibid.*
27. *C-252/21 Meta Platforms Inc and Others v Bundeskartellamt* ECLI:EU:C:2023:537 [73].
28. *Valsts policijas Rīgas reģiona pārvaldes Kārtības policijas pārvalde v Rīgas pašvaldības SIA "Rīgas satiksme* [2017] 4 WLR 97 [28]-[32].
29. *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:957 [58].
30. *Ibid.* [59], [66].
31. *PM et al v OpenAI LP* (N.D. Cal. 2023) [274], [311], [490].
32. AG opinion in *Poland v Parliament and Council* [2021] ECLI:EU:C:2021:613 [AG 73].
33. *Ibid.* [AG 71].
34. *C-401/19 Poland v Parliament and Council* [2022] ECLI:EU:C:2022:297 [44].
35. *Ibid.* [45].
36. *Ibid.* [46].
37. *Ahmet Yildirim v Turkey* App no 3111/10 (ECtHR, 18 March 2013) [47], [64].
38. *Poland v Parliament and Council* [2022] ECLI:EU:C:2022:297 [68].
39. AG opinion in *Poland v Parliament and Council* [2021] ECLI:EU:C:2021:613 [AG 90].
40. *Rotaru v Romania* App no 28341/95 (2000) 8 BHRC 449 [52]; *Kennedy v the United Kingdom* App no 26839/05 (2010) 52 EHRR [151]; *Delfi v Estonia* App no 64569/09 (ECtHR, 16 June 2015) [120]-[122]; *Ahmet Yildirim v Turkey* App no 3111/10 (ECtHR, 18 March 2013) [57].
41. *Glukhin v Russia* App no 11519/20 (ECtHR, 4 July 2023).
42. *Kennedy v the United Kingdom* App no 26839/05 (2010) 52EHRR [151]; *Ekimdzhev and other v Bulgaira* App no 70078/12 (ECtHR, 11 January 2022) [408]-[409]; *Ahmet Yildirim v Turkey* App no 3111/10 (ECtHR, 18 March 2013) [57].

43. *Salov v Ukraine* App no 65518/01 (ECtHR, 6 Sept 2005); *Kwiecień v Poland* App no 51744/99 (ECtHR, 9 January 2007); *Lidia Kita v Poland* App no 27710/05 (ECtHR, 22 July 2008).
44. *Big Brother Watch and others v United Kingdom* App nos 58170/13, 62322/14 and 24960/15 (2018) ECHR 299 [305], [313]; *Rotaru v Romania* App no 28341/95 (2000) 8 BHRC 449 [52]; *S and Marper v the United Kingdom* (2009) 48 EHRR 50 [95].
45. C-673/17 *Bundesverband der Verbraucherzentralen und Verbraucherverbände – Verbraucherzentrale Bundesverband e.V. v Planet49 GmbH* [2019] EU:C:2019:246 [74], [75].
46. *Węgrzynowski and Smolczewski v Poland* App No 33846/07 (ECtHR, 16 July 2007) [58].
47. *Big Brother Watch and others v United Kingdom* App nos 58170/13, 62322/14 and 24960/15 (2018) ECHR 299 [204]; *Ahmet Yildirim v Turkey* App no 3111/10 (ECtHR, 18 March 2013) [57].
48. *Ahmet Yildirim v Turkey* App no 3111/10 (ECtHR, 18 March 2013); *Delfi v Estonia* App no 64569/09 (ECtHR, 16 June 2015).
49. AG opinion in C-70/10 *Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM)* [2011] ECLI:EU:C:2011:255 [AG 53]–[AG 59]; *Tele2 Sverige AB v Postoch telestyrelsen* [2016] All ER (D) 107 (Dec) and *Secretary of State for the Home Department v Tom Watson* [2016] All ER (D) 107 (Dec) [121].
50. *Poland v Parliament and Council* [2022] ECLI:EU:C:2022:297 [24], [90].
51. C-18/18 *Eva Glawischnig-Piesczek v Facebook Ireland Limited* [2019] ECLI:EU:C:2019:821 [41]–[46].
52. *Wikimedia Foundation Inc v Turkey* App no 25479/19 (ECtHR, 24 March 2022) [19]; *Cengiz and others v Turkey* App nos 48226/10 and 14027/11 (ECtHR, 1 December 2015) [29].
53. *Barbulescu v Romania* App no 61496/08 (ECtHR, 5 September 2017) [110], [122]; *Klass and others v Germany* App no 5029/71 (1979–1980) 2 EHRR 214 [55].
54. *Big Brother Watch and others v United Kingdom* App nos 58170/13, 62322/14 and 24960/15 (2018) ECHR 299 [318]; *Rotaru v Romania* App no 28341/95 (2000) 8 BHRC 449 [59], [122].
55. *Tele2 Sverige AB v Postoch telestyrelsen* [2016] All ER (D) 107 (Dec) and *Secretary of State for the Home Department v Tom Watson* [2016] All ER (D) 107 (Dec) [123]; C-40/17 *Fashion ID GmbH & Co.KG v Verbraucherzentrale NRW eV* [2019] [17].
56. *M et al v OpenAI LP.*, 3:23-cv-03199 (US District Court, N.D. Cal. 2023) [152], [248], [249], [269], [298].
57. *Ibid.*
58. AG’s Opinion in C-275/06 *Productores de Musica de Espana (Promusicae) v Telefonica de Espana SAU* [2008] ECR I-271 [AG 121].
59. *Tele2 Sverige AB v Postoch telestyrelsen* [2016] All ER (D) 107 (Dec) and *Secretary of State for the Home Department v Tom Watson* [2016] All ER (D) 107 (Dec) [123].
60. C-460/20 *TU, RE v Google* [2022] ECLI:EU:C:2022:962; C-136/17 *GC, AF, BH, ED v Commission nationale de l’informatique et des libertés (CNIL)* [2019] ECLI:EU:C:2019:773; C-507/17 *Google LLC, successor in law to Google Inc. v Commission nationale de l’informatique et des libertés (CNIL)* [2019] ECLI:EU:C:2019:772; C-131/12 *Google Spain SL Google Inc. v Agencia Española de Protección de Datos and Mario CostejaGonzález* [2014] ECLI:EU:C:2014:317.
61. See e.g. *S and Marper v the United Kingdom* App nos 30562/04 and 30566/04 (2008) ECHR 1581 [119], [124], [125].
62. *PM et al v OpenAI LP* (N.D. Cal. 2023) [311].
63. *Ibid.* [274] and [490].
64. *Kurić and Others v Slovenia* App no 26828/06 (ECtHR, 26 June 2012) [348], [349].
65. *Delfi v Estonia* App no 64569/09 (ECtHR, 16 June 2015) [78]; *Khurshid Mustafa and Tarzibachi v Sweden* App no 23883/06 (ECtHR, 16 March 2009) [43].
66. *Glukhin v Russia* App no 11519/20 (ECtHR, 4 July 2023) [55], [84].
67. *S and Marper v the United Kingdom* App no 30562/04 and 30566/04 (2008) ECHR 1581 [101]; *Peck v the United Kingdom* App no 44647/98 (2003) 36 EHRR 41 [76]; *Khurshid Mustafa and Tarzibachi v Sweden* App no 23883/06 (ECtHR, 16 March 2009) [42].
68. *Ibid.*
69. *Delfi v Estonia* App no 64569/09 (ECtHR, 16 June 2015) [78].

70. *Barbulescu v Romania* App no 61496/08 (ECtHR, 5 September 2017) [121]; *James and Others v the United Kingdom* App no 8793/79 (ECtHR, 21 February 1986) [51]; *Ahmet Yildirim v Turkey* App no 3111/ 10 (ECtHR, 18 March 2013) [64].
71. *Ibid.*
72. C-287/11 *Aalberts Industries NV and Others v European Commission* [2013] EUECJ [54]-[57]; C443/13 *Ute Reindl v Bezirkshauptmannschaft Innsbruck* [2014] EUECJ [39].
73. AG opinion in *Poland v Parliament and Council* [2021] ECLI:EU:C:2021:613 [214].
74. *Barbulescu v Romania* App no 61496/08 (ECtHR, 5 September 2017) [110], [122]; *Rotaru v Romania* App no 28341/95 (2000) 8 BHRC 449 [59].
75. *Taylor-Sabori v the United Kingdom* App no 47114/99 (ECtHR, 12 October 2002) [17]-[19]; *Radu v the Republic of Moldova* App no 50073/07 (ECtHR, 15 April 2014) [31].
76. C-291/12 *Michael Schwarz v Stadt Bochum* [2013] ECLI:EU:C:2013:670; C-131/12 *Google Spain SL Google Inc. v Agencia Española de Protección de Datos and Mario Costeja González* [2014] ECLI:EU:C:2014:317.
77. C-184/20 *OT v Vyriausioji tarnybinės etikos komisija* [2022] EUECJ [127].
78. C-252/21 *Meta Platforms Inc and Others v Bundeskartellamt* [2023] EUECJ [73].
79. *PM et al v OpenAI LP* (N.D. Cal. 2023) [217].
80. AG opinion in C-634/21 *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:220 [AG 58].
81. *Ibid.*
82. *PM et al v OpenAI LP* (N.D. Cal. 2023) [151], [202], [228].
83. *Ibid.* [229], [236].
84. *Glukhin v Russia* App no 11519/20 (ECtHR, 4 July 2023) [56], [57]; *Delfi v Estonia* App no 64569/09 (ECtHR, 16 June 2015) [159].
85. *Magyar Tartalomszolgáltatók Egyesülete and Index.hu Zrt v Hungary* App no 22947/13 (ECtHR, 2 January 2016) [58], [59]; *Pihl v Sweden* App no. 74742/14 (ECtHR, 9 March 2017) [26], [29].
86. C-360/10 *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV* [2012] ECLI:EU:C:2012:85 [47]-[51]; C-70/10 *Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM)* [2012] ECLI:EU:C:2011:771 [53].
87. *Poland v Parliament and Council* [2022] ECLI:EU:C:2022:297 [75].
88. *PM et al v OpenAI LP* (N.D. Cal. 2023) [155].
89. *Ibid.* [247], [258].
90. *Ibid.* [151], [163].
91. *Ibid.* [219].
92. *Glukhin v Russia* App no 11519/20 (ECtHR, 4 July 2023) [77].
93. *Ibid.* [83].
94. *Ibid.* [85], [86].
95. *Ibid.* [89], [90], [91]. In this context, see also Romero-Moreno, F. 2022. 'Facial recognition technology: how it's being used in Ukraine and why it's still so controversial.' <https://theconversation.com/facial-recognition-technology-how-its-being-used-in-ukraine-and-why-its-still-so-controversial-183171>
96. C-275/06 *Productores de Musica de Espana (Promusicae) v Telefonica de Espana SAU* [2008] ECR I-271 [68].
97. *Poland v Parliament and Council* [2022] ECLI:EU:C:2022:297 [86], [90].
98. *Ibid.* [99].
99. *PM et al v OpenAI LP* (N.D. Cal. 2023).

Acknowledgements

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. This paper is a heartfelt tribute to Stephen M. Baker, my dear friend, mentor, and second father. His unwavering encouragement and support fostered my curiosity and transformed me into the researcher I am today. This work, exploring deepfakes, directly reflects his passion for technology and civil liberties. And though his absence leaves a void, his spirit of laughter, intellectual pursuit and unwavering support will forever live on.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

- Ajder, H., and J. Glick. 2021. "JUST JOKING! Deepfakes, Satire and the Politics of Synthetic Media." <https://cocreationstudio.mit.edu/just-joking/>.
- Barrington, S., Romit Barua, Gautham Koorma, and Hany Farid. 2023. "Single and Multi-Speaker Cloned Voice Detection: From Perceptual to Learned Features." ArXiv. <https://arxiv.org/abs/2307.07683>.
- Biometric Update. 2024. "Deepfake Videos Looked So Real that an Employee Agreed to Send Them \$25 Million." <https://www.biometricupdate.com/202402/deepfake-videos-looked-so-real-that-an-employee-agreed-to-send-them-25-million>.
- Brookings. 2024. "Detecting AI Fingerprints: A Guide to Watermarking and Beyond." <https://www.brookings.edu/articles/detecting-ai-fingerprints-a-guide-to-watermarking-and-beyond/>.
- Brown, H., Katherine Lee, Fatemehsadat Miresghallah, Reza Shokri, and Florian Tramèr. 2022. "What Does It Mean for a Language Model to Preserve Privacy?" 2280–2292. <https://doi.org/10.1145/3531146.3534642>.
- CEDPO (Confederation of European Data Protection Organisations). 2023. "Generative AI: The Data Protection Implications." <https://cedpo.eu/generative-ai-the-data-protection-implications/>.
- CFR (Council of Foreign Relations). 2018. "Disinformation on Steroids: The Threat of Deep Fakes." <https://www.cfr.org/report/deep-fake-disinformation-steroids#:~:text=A%20well%2Dtimed%20and%20thoughtfully,political%20divisions%20in%20a%20society>.
- Coalition for Content Provenance and Authenticity. 2024. "Introducing Content Credentials Icon." <https://c2pa.org/post/contentcredentials/>.
- Defense One. 2019. "The Newest AI-Enabled Weapon: 'Deep-Faking' Photos of the Earth." <https://www.defenseone.com/technology/2019/03/next-phase-ai-deep-faking-whole-world-and-china-ahead/155944/>.
- EDPS (European Data Protection Supervisor). 2010. "Opinion of the European Data Protection Supervisor on the Current Negotiations by the European Union of an Anti-Counterfeiting Trade Agreement (ACTA)." https://www.edps.europa.eu/data-protection/our-work/publications/opinions/anti-counterfeiting-trade-agreement-acta-0_en.
- EDPS (European Data Protection Supervisor). 2023. "Deepfake Detection." https://www.edps.europa.eu/data-protection/technology-monitoring/techsonar/deepfake-detection_en.
- Edwards, L. 2022. "The EU AI Act: A Summary of its Significance and Scope." <https://www.adalovelaceinstitute.org/wp-content/uploads/2022/04/Expert-explainer-The-EU-AI-Act-11-April-2022.pdf>.
- EP (European Parliament). 2007. "Better Regulation and the Improvement of EU Regulatory Environment, Institutional and Legal Implications of the Use of 'Soft Law' Instruments." [https://www.europarl.europa.eu/RegData/etudes/note/join/2007/378290/IPOL-JURI_NT\(2007\)378290_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/note/join/2007/378290/IPOL-JURI_NT(2007)378290_EN.pdf).
- EP (European Parliament). 2021. "Tackling Deepfakes in European Policy." [https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU\(2021\)690039_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/STUD/2021/690039/EPRS_STU(2021)690039_EN.pdf).
- Equality Now. 2024. "Briefing Paper: Deepfake Image-Based Sexual Abuse, Tech-Facilitated Sexual Exploitation and the Law." <https://equalitynow.storage.googleapis.com/wp-content/uploads/2024/01/17084238/EN-AUDRI-Briefing-paper-deepfake-06.pdf>.
- German Federal Council. 2021. "Proposal for a Regulation of the European Parliament and the Council Laying Down Harmonised Rules on Artificial Intelligence and Amending Certain Union Legislative Acts on AI and Data Protection." https://www.bundesrat.de/SharedDocs/drucksachen/2021/0401-0500/488-21.pdf?__blob=publicationFile&v=1.
- Google. 2023. "Updates to Political Content Policy (September 2023)." <https://support.google.com/adspolicy/answer/13755910?hl=en>.
- Google Cloud. 2023. "Carbon Free Energy for Google Cloud Regions." <https://cloud.google.com/sustainability/region-carbon>.

- Hasan, H., and K. Salah. 2019. "Combating Deepfake Videos Using Blockchain and Smart Contracts." IEEE. <https://ieeexplore.ieee.org/document/8668407>.
- Home Security Heroes. 2023. "2023 State of Deepfakes: Realities, Threats and Impact." <https://www.homesecurityheroes.com/state-of-deepfakes/>.
- ICO (Information Commissioner Office). 2023. "Guidance on AI and Data Protection." <https://ico.org.uk/for-organisations/uk-gdpr-guidance-and-resources/artificial-intelligence/guidance-on-ai-and-data-protection/>.
- Intel. 2022. "Intel Introduces Real-Time Deepfake Detector." <https://www.intel.com/content/www/us/en/newsroom/news/intel-introduces-real-time-deepfake-detector.html>.
- Intel. 2023. "Generate Structured Synthetic Data: Numeric, Categorical, and Time-Series Tabular Data." <https://www.intel.com/content/www/us/en/developer/articles/reference-kit/ai-structured-data-generation.html>.
- Intelligencer. 2019. "Can You Spot a Deepfake? Does It Matter?" <http://nymag.com/intelligencer/2019/06/how-do-you-spot-a-deepfake-it-might-not-matter.html>.
- Internet Watch Foundation. 2023. "How AI is Being Abused to Create Sexual Abuse Imagery." <https://www.iwf.org.uk/about-us/why-we-exist/our-research/how-ai-is-being-abused-to-create-child-sexual-abuse-imagery/>.
- Keese, N., and M. R. Leiser. 2024. "Online Manipulation as a Potential Interference with the Right to Freedom of Thought." In *The Cambridge Handbook on the Right to Freedom of Thought*, edited by Bethany Shiner and Patrick O'Callaghan, CUP. Forthcoming. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4703015.
- Kop, M. 2021. "EU Artificial Intelligence Act: The European Approach to AI." <https://law.stanford.edu/publications/eu-artificial-intelligence-act-the-european-approach-to-ai/>.
- Labuz, M. 2023. "Regulating Deep Fakes in the Artificial Intelligence Act." Applied Cybersecurity and Internet Governance. <https://www.acigjournal.com/Regulating-Deep-Fakes-in-the-Artificial-Intelligence-Act,184302,0,2.html>.
- Leiser, M. R. 2023. "Psychological Patters and Article 5 of the AI Act Proposal." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4631535.
- Mirsky, Y., and W. Lee. 2020. "The Creation and Detection of Deepfakes: A Survey." ACM Computing Services. <https://arxiv.org/pdf/2004.11138.pdf>.
- The MIT Center for Advanced Virtuality. 2021. "Media Literacy in the Age of the Deepfakes." <https://deepfakes.virtuality.mit.edu/>.
- National Cyber Security Centre Annual Review 2023. 2023. "Case Study: Defending Our Democracy in a New Digital Age – at the Ballot Box and Beyond." <https://www.ncsc.gov.uk/collection/annual-review-2023/resilience/case-study-defending-democracy#:~:text=The%20government's%20Defending%20Democracy%20Taskforce,drives%20the%20government's%20election%20preparedness>.
- Novelli, Claudio, Federico Casolari, Philipp Hacker, Giorgio Spedicato, and Luciano Floridi. 2024. "Generative AI in EU Law: Liability, Privacy, Intellectual Property, and Cybersecurity." ArXiv. <https://arxiv.org/abs/2401.07348>.
- OpenAI. 2023. "GPT-4 Technical Report." <https://cdn.openai.com/papers/gpt-4.pdf>.
- Patterson, D., Joseph Gonzalez, Quoc Le, Chen Liang, Lluís-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. 2021. "Carbon Emissions and Large Neural Network Training." ArXiv. <https://arxiv.org/ftp/arxiv/papers/2104/2104.10350.pdf>.
- Romero-Moreno, F. 2019. "'Notice and Staydown' and Social Media: Amending Article 17 of the Proposed Directive on Copyright." *International Review of Law, Computers and Technology*. <https://www.tandfonline.com/doi/full/10.108013600869.2018.1475906>.
- Romero-Moreno, F. 2020. "'Upload Filters' and Human Rights: Implementing Article 17 of the Directive on Copyright in the Digital Single Market." *International Review of Law, Computers and Technology*. <https://www.tandfonline.com/doi/full/10.108013600869.2020.1733760>.
- Saenko, K. 2023. "Is Generative AI Bad for the Environment? A Computer Scientist Explains the Carbon Footprint of ChatGPT and its Cousins." The Conversation. <https://theconversation.com/is-generative-ai-bad-for-the-environment-a-computer-scientist-explains-the-carbon-footprint-of-chatgpt-and-its-cousins-204096>.
- Sensity. 2023. "Deepfake Detection." <https://sensity.ai/deepfake-detection/>.

- Shattock, E. 2022. "Fake News in Strasbourg: Electoral Disinformation and Freedom of Expression in the European Court of Human Rights (ECtHR)." *European Journal of Law and Technology*. <https://ejlt.org/index.php/ejlt/article/view/882>.
- Syntheticus. 2023. "The Benefits and Limitations of Generating Synthetic Data." <https://syntheticus.ai/blog/the-benefits-and-limitations-of-generating-synthetic-data#:~:text=By%20using%20synthetic%20data%2C%20organizations,too%20expensive%20or%20time%2Dconsuming>.
- Thiel, D. 2023. *Identifying and Eliminating CSAM in Generative AI ML Training Data and Models*. Stanford Internet Observatory Cyber Policy Center. https://stacks.stanford.edu/file/druid:kh752sm9123/ml_training_data_csam_report-2023-12-23.pdf.
- Thi Nguyena, T., Quoc Viet Hung Nguyenb, Dung Tien Nguyena, Duc Thanh Nguyena, Thien Huynh-Thec, Saeid Nahavandid, Thanh TamNguyene, Quoc-Viet Phamf, and Cuong M. Nguyeng. 2022. "Deeplearning for Deepfakes Creation and Detection: A Survey." <https://www.sciencedirect.com/science/article/abs/pii/S1077314222001114>.
- United Nations. 2023. "Policy Brief 8 Information Integrity on Digital Platforms." <https://www.un.org/sites/un2.un.org/files/our-common-agenda-policy-brief-information-integrity-en.pdf>.
- Usercentrics. 2023. "What is Global Privacy Control." <https://usercentrics.com/knowledge-hub/what-is-global-privacy-control/>.
- Van der Sloot, B., and Y. Wagensveld. 2022. "Deepfakes: Regulatory Challenges for the Synthetic Society." *Computer Law and Security Review*. <https://www.sciencedirect.com/science/article/pii/S0267364922000632>.
- Veale, M., and F. Z. Borgesius. 2021. "Demystifying the Draft EU Artificial Intelligence Act." *Computer Law Review International*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3896852.
- Wahlster, W., and C. Winterhalter. 2022. "German Standardization Roadmap on Artificial Intelligence." <https://www.din.de/resource/blob/916798/ed09ae58b60f0d3a498fa90fa5085b7c/nrm-ki-engl-2023-final-web-250-neu-data.pdf>.