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

Preprint · September 2023

DOI: 10.13140/RG.2.2.20397.05604

1 author:

Felipe Romero-Moreno
University of Hertfordshire
14 PUBLICATIONS **121** CITATIONS

SEE PROFILE

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

Keywords: generative AI, deepfakes, human rights

Abstract

This paper critically assesses to what extent under the EU AI Act (Act), the provisions governing the use of deepfakes could be implemented in a way which is compatible with the right of AI providers and users to privacy and freedom of expression under Articles 8 and 10 of the European Convention on Human Rights (ECHR), and the General Data Protection Regulation (EU) 2016/679 (GDPR). The analysis draws on the Act deepfake provisions, the case-law of the Strasbourg and Luxembourg courts, and academic literature. It critically examines the compatibility of the deepfake provisions with the European Court of Human Rights' (ECtHR) three-part, non-cumulative test to determine whether the obligations set out in the Act concerning AI providers and users can be adopted: firstly, that it is 'in accordance with the law'; secondly, that it pursues one or more legitimate aims included in Article 8(2) and 10(2) Convention; and thirdly, that it is 'necessary' and 'proportionate'. The paper addresses a significant gap in the literature. It proposes that the Act be amended to introduce new obligations for AI providers oblige them to deploy structured synthetic data to detect deepfakes, and in addition to electoral disinformation, also explicitly consider AI systems intended to be used for sextorsion and AI-child pornography, high-risk AI. It concludes that unless, following Article 7(1), empowering the Commission to amend the Act, the proposals in the paper for procedural safeguards are implemented, its deepfake provisions will violate Articles 8 and 10 ECHR, and the GDPR.

1. Introduction

John McCarthy coined the term artificial intelligence (AI) in 1955. There are multiple definitions of AI, but a general one is 'the capability of a machine to imitate intelligent human behaviour' (Ofcom, 14).[1] AI has different levels with increasing versatility and sophistication. Firstly, Artificial Narrow Intelligence, can perform a specific task or set of tasks but it cannot generalize to other tasks, for instance, generative AI large language models creating complex content like text, images, video or audio. Moreover, Artificial General Intelligence would understand or learn any intellectual

[1] Ofcom. 2019. "Use of AI in Online Content Moderation." https://www.ofcom.org.uk/__data/assets/pdf_file/0028/157249/cambridge-consultants-ai-content-moderation.pdf.

task that a human can. Additionally, Artificial Super Intelligence would surpass human intelligence in every way (Ofcom, 14).[2]

Recital 6a of the EU Artificial Intelligence Act (hereinafter the Act) explains that AI can be trained using machine learning to perform new tasks without being explicitly programmed. Machine learning techniques include supervised learning, unsupervised learning, and reinforcement learning, and use methods such as deep learning with neural networks. Article 3(1c) defines foundation models as AI systems trained on broad large-scale data that can be adapted to a variety of different tasks and generate general outputs. Moreover, under Article 3(1d), general purpose AI means systems which can be adapted to a variety of applications and used for unintended purposes.

In 2020, OpenAI launched GPT-3, a large language model trained on extensive datasets which is versatile in natural language processing tasks like question answering, summarisation and translation (OpenAI 2021).[3] Based on GPT-3, in December 2022, OpenAI released its generative AI application ChatGPT (OpenAI 2022).[4] OpenAI introduced GPT-4 (OpenAI 2023),[5] in March 2023. It further expanded the beneficial applications of generative AI to art generation, coding and creative writing. Google is also developing generative AI, including the conversational chatbot Bard (Google 2023).[6] In March 2023, Google made its generative AI tools (Google Cloud 2023)[7] available to developers, businesses and governments, which can be used to create content, automate tasks, and build applications (EP 2023, 1-2).[8]

However, as the US class action complaint *Clarkson v OpenAI* warned, despite its useful applications, generative AI also posed risks, such as facilitating the widely rapid spread of 'deepfakes', to disseminate misinformation, extort victims, or access classified or confidential information.[9] The challenge is that the growing simplicity of creating realistic deepfakes and fake media environments is no longer limited to experts. Indeed, not just can a deepfake video be created, but equally fake X accounts which post links to the video, fake social media accounts that comment on the video, fake sites which host the YouTube video creating deliberately misleading

---

[2] *Ibid*.
[3] OpenAI. 2021. "GPT-3 powers the next generation of apps." https://openai.com/blog/gpt-3-apps.
[4] OpenAI. 2022. "Introducing ChatGPT." https://openai.com/blog/chatgpt.
[5] OpenAI. 2023. "GPT-4 is OpenAI's most advanced system, producing safer and more useful responses." https://openai.com/gpt-4.
[6] Google Bard experiment. 2023. https://bard.google.com/.
[7] Google Cloud. 2023. "Google Cloud brings generative AI to developers, businesses, and governments." https://cloud.google.com/blog/products/ai-machine-learning/generative-ai-for-businesses-and-governments.
[8] EP (European Parliament). 2023. "General-purpose artificial intelligence." https://www.europarl.europa.eu/RegData/etudes/ATAG/2023/745708/EPRS_ATA(2023)745708_EN.pdf.
[9] *PM et al v OpenAI LP.*, 3:23-cv-03199 (US District Court, N.D. Cal. 2023) [219].

or false information, fake Instagram accounts that create memes using the video, etc (Van der Sloot and Wagensveld 2022, 2).[10] Worryingly, this makes it easier for deepfakes to be used to deceive or manipulate people, but harder for both humans and AI to track and verify information.

Before making a generative AI model publicly available or deploying it, Article 28 of the Act requires providers to satisfy the transparency obligations in Article 52(1), also design, train, and develop the model adopting appropriate safeguards, as well as documenting and reporting on its training data use, protected under copyright law. Conversely, pursuant to Article 52(3)1, users of AI systems which create or manipulate content are required to disclose in a timely, visible and clear way that the content is deepfake, and if possible, the natural or legal person name that created or manipulated it. However, Article 52(3)2 clarifies that such disclosure obligation shall not apply if the use of the AI systems is authorised by law or if it is necessary to enjoy the right to freedom of expression and the right to freedom of the arts and sciences under the EU Charter of Fundamental Rights.

Expanding on the author's previous work (Romero-Moreno 2019),[11] and (Romero-Moreno 2020),[12] this research has two main goals. The first is to critically examine to what extent under the EU AI Act, the provisions governing the use of deepfakes could be implemented in a way which is compatible with the right of AI providers and users to privacy and freedom of expression under Articles 8 and 10 of the European Convention on Human Rights (ECHR), and the General Data Protection Regulation (EU) 2016/679 (GDPR). The second is to propose and evaluate some procedural safeguards to make the Act deepfake provisions compliant with the ECHR, and the GDPR. The paper addresses a significant gap in the literature. It suggests that the Act be amended to include new provisions for AI providers requiring them to use structured synthetic data to detect deepfakes, and in addition to electoral disinformation, also expressly consider AI meant to be deployed for sextorsion and AI-child pornography, high-risk AI. I conclude that unless, following Article 7(1), empowering the Commission to amend the Act, the procedural safeguards recommended in the paper are adopted, its provisions governing deepfakes will violate Articles 8 and 10 of the Convention, and the GDPR.

---

[10] Van der Sloot, B. and Wagensveld, Y. 2022."Deepfakes: regulatory challenges for the synthetic society." *Computer Law and Security Review*. https://www.sciencedirect.com/science/article/pii/S0267364922000632.
[11] 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/10.1080/13600869.2018.1475906.
[12] Romero-Moreno, F. 2020. "'Upload filters' and human rights: implementing Article 17 of the EU Copyright Directive in the Digital Single Market." *International Review of Law, Computers and Technology*. https://www.tandfonline.com/doi/full/10.1080/13600869.2020.1733760.

2. Defining deepfakes

The term ‘deepfake’ is a combination of ‘deep learning’ and ‘fake’. It was first coined by an anonymous user on Reddit in 2017. The user created a subreddit called ‘deepfakes’ where videos were posted, which used AI to swap the faces of celebrities with those of other people, a process known as face swapping (Thi Nguyen et al. 2022, 1).[13] The first videos contained pornographic content, where the faces of the original actresses were replaced by those of Scarlett Johansson, Taylor Swift and Gal Gadot. The user also shared the programming code, which enabled others to create their own deepfakes (The Guardian News, January 13, 2020). Article 3(1) point 44d of the Act defines deepfakes as synthetic or manipulated image, audio or video content, which would deceptively seem to be truthful or authentic, and that contains portraits of individuals appearing to do or say things they did not do or say, created using AI-based machine learning and deep learning.

3. Beneficial use of deepfakes

Deepfakes can be used for positive purposes, such as in visual effects, Snapchat filters, digital avatars, generating synthetic voices for individuals who have lost theirs, or updating existing movie episodes without reshooting them.[14] Moreover, other productive and creative uses of deepfakes include gaming, virtual reality, photography, and entertainment. For example, deepfakes can be used to allow people to virtually try on clothes before they buy them, make foreign films more accessible to a wider audience, or educate people through reanimated historical figures (Mirsky and Lee 2020, 1).[15] Worryingly, however, while deepfakes can be used for positive purposes, the number of malicious uses is much greater.

4. Harmful use of deepfakes

Given the amount of data available online for public figures such as celebrities and politicians, unsurprisingly, deepfakes are often used to target these individuals. However, it puts international security at risk when generative AI can also be used to create deepfakes of world leaders giving speeches that are not genuine,[16] or fake

---

[13] Thi Nguyena, T., Quoc Viet Hung Nguyenb, Dung Tien Nguyena, Duc Thanh Nguyena, Thien Huynh-Thec, Saeid Nahavandid, Thanh Tam Nguyene, Quoc-Viet Phamf, Cuong M. Nguyeng. 2022. “Deeplearning for deepfakes creation and detection: a survey.” *Computer Vision and Image Understanding*. https://www.sciencedirect.com/science/article/abs/pii/S1077314222001114.
[14] Forbes. 2019. “The best (and scariest) examples of AI-enabled deepfakes.” https://www.forbes.com/sites/bernardmarr/2019/07/22/the-best-and-scariest-examples-of-ai-enabled-deepfakes/.
[15] Mirsky, Y., and Lee, W. 2020. “The creation and detection of deepfakes: a survey.” *ACM Computing Services*. https://arxiv.org/pdf/2004.11138.pdf
[16] Bloomberg. 2018. “How faking videos became easy and why that’s so scary.” https://www.bloomberg.com/news/articles/2018-09-10/how-faking-videos-became-easy-and-why-that-s-so-scary-quicktake?leadSource=uverify%20wall.

satellite images, which include objects that do not really exist.[17] Yet, even without being shared on social media, the intelligence services could also use deepfakes to target presidential advisors to attempt to control global decision-making.[18]

Recital 40a of the Act states that AI intended to be deployed to influence elections, referendum results or individual behaviour should be considered high-risk AI. In this context, it is significant that *Clarkson v OpenAI* recognised how deepfakes could also meddle in elections, sow distrust, and jeopardise public discourse. Indeed, based on US Congressional Research Service's evidence, the class action cautioned about the risks of deepfakes used not just to extort officials or individuals with access to privileged information, but also create content to enlist terrorists, radicalise groups or promote violence.[19]

It should be noted however that, deepfakes can harm everyone, regardless of their position or status. For example, acknowledging the FBI warning, *Clarkson v OpenAI* also alerted about the threat of 'sextortion', being perpetrated using generative AI and public domain images and videos of ordinary individuals through social media to create deepfake pornography. This content, which alarmingly, involved not only non-consenting adults, but also kids, was then shared widely on pornographic sites, and public forums to harass the victim, inflicting both significant psychological and emotional harm.[20] 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.[21]

Furthermore, *Clarkson v OpenAI* also drew attention to how OpenAI's Dall-E model was often used by pedophiles, because it was less technically demanding than previous systems and enabled the production of images of virtual child pornography at a larger scale.[22] Dall-E was trained on a massive dataset of images collected, without knowledge or permission, many of which displayed real kids and was considered the source material for AI-child pornography. Disturbingly, sometimes real images of existing child pornography were deployed to train the model and

---

[17] 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/.
[18] CFR. 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.
[19] *PM et al v OpenAI LP.*, 3:23-cv-03199 (US District Court, N.D. Cal. 2023) [220].
[20] *Ibid*. [222], [223].
[21] *Ibid*. [223].
[22] *Ibid*. [224].

created further explicit content of previously abused kids, thereby re-traumatising them and exacerbating their pain.[23]

Deepfakes can additionally 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. However, unfortunately the opportunity was a scam, not a legitimate investment.[24] Moreover, 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,[25] to enhance their appearance to prey on people's vulnerabilities (Los Angeles Times, May 11, 2023).

5. Creating deepakes

Deepfakes can be created using several techniques, which differ depending on the source of the content being manipulated. Firstly, video deepfakes is a technique based on neural networks, which uses an autoencoder to identify the salient features of the source video, such as the person's body language and facial expressions. The autoencoder includes an encoder to extract such features and a decoder to place them onto the target video (Github Deepfakes Faceswap).[26] Moreover, a second technique is to create audio deepfakes, where a generative adversarial network (GAN), learns the vocal patterns of a person's voice and then creates a model of such person's voice to speak anything desired (Github Faceswap-GAN).[27] Additionally, based on recurrent neural networks, lip-syncing is another technique, which matches the movement of a person's lips to a pre-recorded audio, making it appear that the person is speaking the recording words. If the audio is also a deepfake, this method makes the deception even more believable (Agarwal et al. 2019).[28]

---

[23] *Ibid*. [225].
[24] MoneySavingExpert. 2023. "Warning: beware terryfinying new 'deepfake' Martin Lewis video scam promoting a fake 'Elon Musk investment' - it's not real." https://www.moneysavingexpert.com/news/2023/07/beware-terrifying-new--deepfake--martin-lewis-video-scam-promoti/.
[25] Midjourney. 2023. https://www.midjourney.com/home/?callbackUrl=%2Fapp%2F.
[26] Github Deepfakes Faceswap. "Deepfakes software for all." https://github.com/deepfakes/faceswap; see also Github DeepFaceLab. "DeepFaceLab is the the leading software for creating deepfakes." https://github.com/iperov/DeepFaceLab; Github Dfaker. "Larger resolution face masker, weirdly warped, deepfake." https://github.com/dfaker/df; Github StrongWine Deepfake_tf. "Deepfake based on tensorflow." https://github.com/StromWine/DeepFake_tf.
[27] Github Faceswap-GAN. "A denoising autoencoder + adversarial losses and attention mechanisms for face swapping." https://github.com/shaoanlu/faceswap-GAN; see also Github CycleGAN. "Image-to-image translation in pytorch." https://github.com/junyanz/pytorch-CycleGAN-and-pix2pix.
[28] Agarwal, S., Hany Farid, Yuming Gu, Mingming He, Koki Nagano, and Hao Li. "Protecting world leaders against deep fakes." *The Computer Vision Foundation*. https://openaccess.thecvf.com/content_CVPRW_2019/papers/Media%20Forensics/Agarwal_Protecting_World_Leaders_Against_Deep_Fakes_CVPRW_2019_paper.pdf.

6. Detecting deepfakes

In addition to deepfakes creation techniques, there are also methods of deepfakes 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.[29] 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.[30]

It is noteworthy that, while Sensity claims that it can identify with 95.8 percent accuracy realistic full bodies and faces generated using AI models like Dall-E,[31] Intel asserts that its technology is the first real-time system, with 96 percent precision.[32] In this context, the CJEU AG opinion in *Poland v Council and Parliament* stressed that, if it were not feasible, to filter content without leading to a significant 'false positive' rate, having a limited effect, such measures should be excluded.[33] Furthermore, in *UPC Telekabel Wien* the CJEU held that, to strike a fair balance, it was crucial under Article 16 Charter, to allow companies to choose the measures they will take, considering their capabilities and resources.[34]

Besides the above Sensity and FakeCatcher systems, there are also other detection tools like the Coalition for Content Provenance Authenticity,[35] Meta AI's Deepfake Detection Challenge Dataset,[36] or DARPA's Media Forensics.[37] However, the difficulty is that *Clarkson v OpenAI* also warned about the overfitting problem, where a torrent of AI-child pornography 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.[38]

---

[29] Sensity. 2023. "Deepfake detection." https://sensity.ai/deepfake-detection/.
[30] 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
[31] Sensity. 2023. "Deepfake detection." https://sensity.ai/deepfake-detection/.
[32] 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.
[33] AG Opinion in C-401/19 *Poland v Parliament and Council* [2021] ECLI:EU:C:2021:613 [214].
[34] C-314/12 *UPC Telekabel Wien GmbH v Constantin FilmVerleih GmbH and Wega Filmproduktionsgesellschaft GmbH* [2013] EU:C:2014:192 [52]; see also C-401/19 *Poland v Parliament and Council* [2022] ECLI:EU:C:2022:297 [75].
[35] Coalition for Content Provenance and Authenticity. 2023. "An open technical standard providing publishers, creators, and consumers the ability to trace the origin of different types of media." https://c2pa.org/.
[36] Meta AI. 2020. "Deepfake Detection Challenge Dataset." https://ai.meta.com/datasets/dfdc/.
[37] Defense Advanced Research Projects Agency. 2019. "Media Forensics (MediFor) (Archived)." https://www.darpa.mil/program/media-forensics.
[38] *PM et al v OpenAI LP.*, 3:23-cv-03199 (US District Court, N.D. Cal. 2023) [226].

7. Deepfakes under the EU AI Act

The EU AI Act adopts a risk-based approach to the regulation of AI. The legal framework includes obligations for providers and users of AI systems based on the potential risks that the AI system can cause. The Commission differentiates between 'unacceptable risk', 'high risk', 'limited risk' and 'minimal risk' (EC 2021).[39]

7.1. Unacceptable risk

Unacceptable risk AI systems are a narrow range of AI applications, which violate human rights. For example, the Act 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 using facial recognition (EC 2021).[40]

7.2. High risk

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 third-party conformity assessment and fall into the specific high-risk areas requiring registration in the EU database (EC 2021).[41]

For instance, Recital 40a of the Act explains that to tackle the risks of unjustified external meddling with the right to vote included in Article 39 of the Charter, AI systems intended to be deployed to influence referendum results, elections or individual voting behaviour should be classified as high-risk AI.

Moreover, Annex III paragraph 1 8 (a b) elaborates that very large social media online platforms as understood in Article 33 of Regulation EU 2022/2065 Digital Services Act (DSA), which deploy AI-based recommendation systems to suggest user-generated content to users of such platforms, should also be considered high-risk AI. The DSA is not a deepfake-specific regulation, but it could be applied to deepfakes, nonetheless, this is not something that this paper will address.

---

[39] EC (European Commission). 2021. “New rules for Artifical Intelligence - Questions and Answers.” htthttps://www.europarl.europa.eu/news/en/headlines/society/20230601STO93804/eu-ai-act-first-regulation-on-artificial-intelligence#:~:text=Generative%20AI%2C%20like%20ChatGPT%2C%20would.copyrighted%20data%20used%20for%20trainingps://ec.europa.eu/commission/presscorner/api/files/document/print/en/qanda_21_1683/QANDA_21_1683_EN.pdf.

[40] *Ibid*.

[41] *Ibid*.

Of particular interest, Recital 60g clarifies that generative AI foundation models should be transparent about the fact that the content is AI-generated and not the product of human creation. However, the specific rules and obligations governing foundation models are not sufficient to consider them high-risk AI systems.

7.3. Limited risk

To mitigate the risk of manipulation, there are also specific transparency requirements in place for certain AI systems, such as those which use generative AI (EC 2021).[42] Article 28b 4a of the Act states that providers of generative AI foundation models shall satisfy the transparency obligations included in Article 52(1), ensuring that AI systems meant to interact with humans are developed, so that the AI system, the provider, or the user tells the human that they are interacting with an AI system in a timely, clear, and understandable way, unless it is clear from the situation and the way the AI system is being used. If applicable, this information shall also contain the enabled AI functions, whether there is human supervision, who is responsible for the decision-making process, as well as the existing procedures and rights enabling individuals or their representatives to object to the use of AI and obtain judicial redress against its decisions, including their right to request an explanation. However, this obligation does not apply to AI systems which are legally allowed to detect, prevent, investigate and prosecute criminal offences, except if such systems are accessible to the public to report a crime.

Furthermore, under Article 28b 4b generative AI foundation model providers shall also design, train, and develop the model in a way, which guarantees appropriate safeguards against the creation of illegal content violating EU law, and without unduly restricting fundamental rights, such as the freedom of expression.

Pursuant to Article 28b 4c, such providers shall also document and make publicly available a sufficiently comprehensive report on the use of training data, which is protected under domestic or EU copyright law.

Importantly, users should also be aware that AI systems can be deployed to create realistic content that can be difficult to distinguish from real one, including deepfakes. Thus, Article 52(3)(1) stresses that users of AI systems which create or manipulate text, image or video content, portraying individuals appearing to do or say things they did not do or say, without their permission, shall disclose in a timely, visible and clear way that the content is deepfake, and if possible, the natural or

[42] *Ibid*.

legal person name that created or manipulated it. Moreover, under Article 52(3)(1), disclosure means labelling the content, so that it is clearly visible for its recipient and informs it is synthetic media.

However, Article 52(3)(2) asserts that paragraph 3 does not apply if the use of deepfakes is allowed by law or if it is necessary to enjoy the right to freedom of expression and the right to freedom of the arts and sciences under the Charter. If the content is artistic, satirical, fictional or creative filmmaking, video games visuals etc, the transparency obligations are restricted to disclosing the deepfake in an adequate and visible clear way, not affecting the content display, and, if appropriate, disclosing the relevant copyrights. Although this shall not preclude law enforcement from deploying AI systems to detect deepfakes and prevent, investigate and prosecute criminal offences.

Given that foundation models are a new and rapidly developing AI field, Recital 60h concludes that it is apt for the Commission and the AI Office to monitor and regularly evaluate the legal and governance framework of generative AI models that lead to significant questions regarding content creation, which infringes EU law, copyright regulations, and likely misuse.

7.4. Minimal Risk

The rest of AI systems can be developed and used in compliance with the existing laws and regulations without being required to meet additional legal requirements. Therefore, most AI systems currently deployed within the EU are part of this category such as, intelligent video games, spam filters and stock control systems (EC 2021).[43]

8. Deepfakes under the GDPR

To create a deepfake, someone typically needs to use data that can be linked to a specific person, such as voice recordings, photos, or videos. If a deepfake portrays a real individual, it can be considered personal data because it can be used to identify the person as it relates to an identified or identifiable natural person (EP 2021, 38).[44] Article 2(5a) Act explains that the processing of personal data is subject to certain conditions under EU data protection and privacy law including the GDPR,

[43] *Ibid*.
[44] 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.

the Law Enforcement Directive, the ePrivacy Directive, and Articles 7 and 8 of the Charter.

The GDPR's definition of processing is broad enough to cover all aspects involved in creating and distributing a deepfake, from gathering the data to deploying it to create the deepfake. The GDPR applies to the development of deepfake applications and software because they use personal data to train their algorithms. This means that deepfake developers process personal data, even if they are not creating specific deepfakes. The GDPR also applies to the creation and dissemination of deepfakes because they also use personal data to create the deepfake content. This indicates that deepfake creators and distributors also process personal data (EP 2021, 39).[45]

The GDPR provides that personal data processing must always have a legal basis. Of the six possible justifications for processing personal data specified in Article 6 GDPR, arguably only consent and legitimate interests could be relevant to deepfake use (EP 2021, 39).[46] Moreover, deepfakes that contain sensitive personal data including people's sexual life (e.g. 'sextorsion' content), or revealing political opinions (e.g. electoral disinformation content), would be prohibited under Article 9(1) GDPR. It appears that besides deepfakes created with explicit consent or as part of a contractual agreement, none of the legal justifications outlined in Article 9(2) GDPR apply to deepfakes (van der Sloot 2020, 4).[47]

The first possible way that personal data can be legally used to create deepfakes is with the consent of the individual(s) in the unmanipulated content, and the consent of the individual(s), appearing in the manipulated deepfake. However, considering the GDPR provisions,[48] and the CJEU rulings in *Planet49*,[49] and *Orange România*,[50] consent must be freely given, specific, informed and unambiguous. This implies that the individual(s) in both, the unmanipulated and manipulated content, must have actively agreed to the personal data processing and that they were given intelligible, easily accessible, and concise information about the potential benefits and risks of giving consent.[51]

---

[45] *Ibid*.
[46] *Ibid*.
[47] Van der Sloot, B. 2020. "Editorial." https://bartvandersloot.com/onewebmedia/edpl_2020_04-004.pdf.
[48] See Recital 32 of GDPR and Article 4(11) GDPR.
[49] C-673/17 *Bundesverband der Verbraucherzentralen und Verbraucherverbände – Verbraucherzentrale Bundesverband e.V. v Planet49 GmbH* [2019] EU:C:2019:246 [61].
[50] C-61/19 *Orange Romania SA v Autoritatea Naţională de Supraveghere a Prelucrării Datelor cu Caracter Personal (ANSPDCP)* [2020] ECLI:EU:C:2020:901 [8].
[51] See C-673/17 *Bundesverband der Verbraucherzentralen und Verbraucherverbände – Verbraucherzentrale Bundesverband e.V. v Planet49 GmbH* [2019] EU:C:2019:246 [61]; and C-61/19 *Orange Romania SA v Autoritatea Naţională de Supraveghere a Prelucrării Datelor cu Caracter Personal (ANSPDCP)* [2020] ECLI:EU:C:2020:901 [39], [40], [52].

However, the challenge is that although deepfakes can be created with individual's explicit consent involving the unmanipulated and/or manipulated content, most deepfakes are created, published, viewed and shared without freely given, specific, informed and unambiguous consent of the individuals being impersonated. Indeed, as noted above, *Clarkson v OpenAI* warned about how 'sextortion' schemes were perpetrated using generative AI to create deepfake pornography. This content which was obtained without knowledge involved not just non-consenting adults, but also kids. It was then made available to the general public to harass the victim, creating significant emotional distress and trauma.[52]

If there is no consent to use personal data for deepfakes purposes, then the legitimate interest basis requires a careful assessment of the potential impact on individual rights and freedoms. When assessing whether deepfake developers, creators and distributors can rely on legitimate interest as a legal justification for deepfake data processing, Article 6(1)(f) GDPR does not explicitly list the factors to consider. However, pursuant to the CJEU *Rīgas satiksme*, deepfake developers, creators and distributors' fundamental rights and freedoms would have to be balanced against those of the impersonated individuals.[53]

As noted above, based on Article 52(3)(2) Act, it is conceivable that in safeguarding the rights of deepfake developers, creators and distributors to freedom of expression and freedom of the arts and sciences, under Articles 11 and 13 of the Charter, they could use deepfakes to express themselves or creatively for artistic, satirical, or fictional purpose. However, unquestionably, if they were also to deploy deepfakes to create content including electoral misinformation, sextorsion or AI-child pornography, the rights of deepfake victims to data protection and privacy under Articles 7 and 8 Charter, would take precedence over their interests.

Recital 45a of the Act stresses that the right to privacy and protection of personal data must be ensured throughout all the stages of the AI. Thus, the EU data protection principles of data minimization and data protection by design and by default, are critical if data processing entails significant risks to individual rights. Recital 45a elaborates that AI providers and users should also adopt organisational and technical measures including encryption and anonymisation, and use technology that allows algorithms to be applied to data without the need to transfer the data between parties or duplicate data unnecessarily.

---

[52] *PM et al v OpenAI LP.*, 3:23-cv-03199 (US District Court, N.D. Cal. 2023) [222], [223].
[53] C-13/16 *Valsts policijas Rīgas reģiona pārvaldes Kārtības policijas pārvalde v Rīgas pašvaldības SIA "Rīgas satiksme* [2017] 4 WLR 97 [28]-[32].

Pursuant to Article 22 GDPR, the impersonated individuals have the right not to be subject to decisions based solely on automated processing, including their profiling, which has a legal effect or a similar significant impact on them. Thus, if generative AI were to be deployed to create sextorsion, AI-child pornography or electoral misinformation content, Article 22(2) GDPR, would protect the victims from AI decision-making*,* unless they explicitly consented to the deepfake processing. Alternatively, there should be a legal justification under the AI Act for doing so (as is currently the case), or it would be required to perform a contract between the deepfake developers, creators and distributors and the victim.

9. Proposal

Article 7(1) of the Act states that the Commission is empowered to implement delegated legislation pursuant to Article 73 to amend Annex III by adding or changing areas or applications of high-risk AI systems if these pose a significant risk of harm to fundamental rights, democracy and the rule of law. Moreover, Article 84(7) explains that the Commission shall submit adequate proposals to amend the Act as needed, considering the latest technological advancements, the impact of AI systems and in light of the evolving nature of the information society.

Given that, this paper proposes that the Act be amended to include new provisions on the use of deepfakes. In the first place, by mandating that AI providers use synthetic data to detect them. Additionally, since along with electoral disinformation, AI systems used for sextorsion and AI-child pornography can clearly pose a significant risk to fundamental rights, also amend Annex III to explicitly include such systems in the list of high-risk AI. Arguably, unless, following Article 7(1), empowering the Commission to amend the Act, the proposals in the paper for procedural safeguards are implemented, its deepfake provisions will violate Articles 8 and 10 ECHR, and the GDPR.

In *Poland v Parliament and Council* the CJEU AG Saugmandsgaard Øe explained that the uploading of content such as, texts, photographs and videos was a part of the right to freedom of expression and information, under Article 11 Charter. However, the AG recognised that if the content involved the artistic expression that the users made available, its uploading also formed a part of the freedom of the arts, protected under Article 13 Charter and Article 10 Convention.[54] Saugmandsgaard Øe observed that the right enshrined in Article 11 Charter, including the right to access and share information, corresponded to that contained in Article 10

[54] AG Opinion in C-401/19 *Poland v Parliament and Council* [2021] ECLI:EU:C:2021:613 [AG 73].

Convention. Thus, the AG recalled that under Article 52(3) Charter, these two rights were identical in meaning and scope. Accordingly, the AG stressed that Article 11 Charter must be interpreted not only considering Article 10 Convention, but also taking into account the case-law of the Strasbourg Court.[55] It is significant that despite not being binding on the CJEU, the Court followed the AG's recommendations.[56] 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'.[57]

10. Assessment of applicability and compliance with Articles 8 and 10 of the ECHR

10.1. 'In accordance with law'

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.[58] The requirement that the interference be based in domestic legislation is not difficult to meet as the Act, which is statutory law, and the ECtHR facial recognition ruling in *Glukhin v Russia*,[59] offer this. However, concerning the second and third criteria, this section argues that the Act could be inconsistent with the Court's accessibility, foreseeability and rule of law principles, thus contravening the first criterion of its three-part, non-cumulative test under Articles 8(2) and 10(2).

In terms of the accessibility principle, it is well-settled Strasbourg Court case-law that legislation must be adequately accessible allowing the individual to have a clear indication of the legal norms, which applied to a given situation.[60] As noted above, Recital 60g of the Act states that generative AI systems should be transparent about

---

[55] *Ibid*. [AG 71].
[56] C-401/19 *Poland v Parliament and Council* [2022] ECLI:EU:C:2022:297 [44], [45], [46], [68].
[57] AG Opinion in C-401/19 *Poland v Parliament and Council* [2021] ECLI:EU:C:2021:613 [AG 90].
[58] *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]; *Liberty and others v the United Kingdom A*pp no 58243/00 (2008) 48 EHRR 1 [59]; *Delfi v Estonia* App no 64569/09 (ECtHR, 16 June 2015) [120]–[122]; *Yildirim v Turkey* App no 3111/10 (ECtHR, 18 March 2013) [57].
[59] *Glukhin v Russia* App no 11519/20 (ECtHR, 4 July 2023).
[60] *Liberty and others v the United Kingdom* App no 58243/00 (2008) 48 EHRR 1 [59]; *Kennedy v the United Kingdom* App no 26839/05 (2010) 52 EHRR [151]; *Ekimdzhiev and other v Bulgaira* App no 70078/12 (ECtHR, 11 January 2022) [408]–[409]; *Yildirim v Turkey* App no 3111/10 (ECtHR, 18 March 2013) [57].

the fact that the AI-generated content is not the product of human creativity, although the specific rules and obligations governing foundation models are not sufficient to consider them high-risk AI. However, regrettably and ignoring ECtHR,[61] and CJEU[62] case-law, the Act fails to address, much less recognise other specific types of deepfake high-risk AI, thereby providers and users being unable to clearly and easily appreciate the impact of such systems. It is true that, prudently, to tackle the risks of unwarranted meddling with the right to vote included in Article 39 EU Charter, Recital 40a of the Act explains that AI meant to be deployed to influence referendum results, elections or individual behaviour should be classified as high-risk AI. Indeed, as noted above, *Clarkson v OpenAI* serves as supporting evidence illustrating how deepfakes could not only influence elections, but also cause political or religious wars worldwide, erode individual trust, and detrimentally affect public discourse.[63] In this context, ECtHR case-law, has also however cautioned that to respect the right to freedom of expression under Article 10 ECHR, law tackling electoral disinformation,[64] should normally only target knowingly false information, which was intended to manipulate voters or erode the rights of others. Conversely, legislation limiting the spread of less deceptive falsehoods such as misinformation,[65] should be evaded, especially if it incorporated criminal sanctions (Shattock 2022, 25).[66] Yet, the Act proposal initially stated that, under Article 52(3), AI models meant to be deployed by law enforcement authorities to detect deepfakes, were considered high-risk AI. Curiously, however, Article 52(3)(2), currently allows law enforcement to deploy those systems to prevent, investigate and prosecute criminal offences without considering the different levels of harm that such systems could inflict. Therefore, one could argue that to satisfy the accessibility principle, in addition to requiring AI providers to undertake conformity assessments of electoral disinformation systems as per Article 43, and register their systems in the EU publicly-accessible database under Article 60, the Act should also consider AI-child pornography and sextorsion high-risk AI. In fact, it should similarly explicitly

---

[61] *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].

[62] C-673/17 *Bundesverband der Verbraucherzentralen und Verbraucherverbände – Verbraucherzentrale Bundesverband e.V. v Planet49 GmbH* [2019] EU:C:2019:246 [74], [75].

[63] *PM et al v OpenAI LP.*, 3:23-cv-03199 (US District Court, N.D. Cal. 2023) [220]; also note that in the US, for example, among other states Texas was the first approving legislation making it a crime to generate and publish deepfakes videos to affect an election outcome recognizing that while this technology cannot be constitutionally prohibited entirely, it can however be specifically targeted to prevent possibly its biggest threat: 'the electoral process'. See Senate Research Center. 2019. "Bill Analysis." https://capitol.texas.gov/tlodocs/86R/analysis/html/SB00751F.htm.

[64] Disinformation involves information that is false and deliberately created to harm a person, social group, organization or country - see page 20 of Wardle, C., and Derakhshan, H. 2017."Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making." https://edoc.coe.int/en/media/7495-information-disorder-toward-an-interdisciplinary-framework-for-research-and-policy-making.html

[65] Misinformation involves information that is false, but not created with the intention of causing harm - see *Ibid*.

[66] 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; see generally for instance *Salov v Ukraine* Application no 65518/01 (ECtHR, 6 Sept 2005); *Kwiecień v Poland* Application no 51744/99 (ECtHR, 9 January 2007); *Lidia Kita v Poland* Application no 27710/05 (ECtHR, 22 July 2008); *Brzeziński v Poland* Application no 47542/07 (ECtHR, 25 Jul 2019); *Jezior v Poland* Application 31955/11 (ECtHR, 4 June 2020); *Staniszweski v Poland* Application no 20422/15 (ECtHR 14 Oct 2021).

mandate that deepfake providers follow the same rules. As the ECtHR succitinly 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.[67] Thus, since other than addressing electoral disinformation, under the Act, providers and users cannot understand how other potentially high-risk AI like sextorsion or AI-child pornography, will affect them, it arguably fails to satisfy the ECtHR accessibility principle under Articles 8(2) and 10(2) ECHR.

As far as the foreseeability principle is concerned, the ECtHR case-law explains that a rule cannot be considered law unless it is laid down precisely enough to allow the individual to reasonably foresee the consequences that an action might involve.[68] As outlined above, Article 52(3)(1) of the Act states that users shall disclose in a timely, visible and clear way that the content is deepfake, and if possible, the natural or legal person name that created or manipulated it. Moreover, under Article 52(3)(1), disclosure means labelling the content, so that it is clearly visible for its recipient and informs it is synthetic media. However, it is concerning that in conflict with ECtHR[69] and CJEU[70] case-law, the Act offers no guidance on whether the platform, which is deployed for communication, should play a role in facilitating and/or monitoring the labelling of deepfakes, for instance, involving electoral disinformation, AI-child pornography or sextorsion. Neither Article 71 spells out among the penalties whether and to which extent user non-adherence to Article 52(3)(1) transparency obligations, is sanctionable (EP 2021, 38).[71] Firstly, it is arguable that to enable users to foresee the consequences of failing to label such synthetic media, the Act should explicitly state that when users publish, view, or share deepfakes, Article 52(3)(1) transparency obligations must not lead to general monitoring obligations imposed on such platforms. In fact, following the CJEU *Poland v Parliament and Council*,[72] and *Glawischnig-Piesczek*,[73] while requiring preventive monitoring of all deepfakes that users wished to upload would lead to prior automatic filtering and

---

[67] *Węgrzynowski and Smolczewski v Poland* App No 33846/07 (ECtHR, 16 July 2007) [58]; see also *Editorial Board of Pravoye Delo and Shtekel v Ukraine* App no 33014/05 (ECtHR, 5 May 2011) [63].
[68] *Big Brother Watch and others v United Kingdom* App nos 58170/13, 62322/14 and 24960/15 (2018) ECHR 299 [204]; *Kennedy v the United Kingdom* App no 26839/05 (2010) 52 EHRR [151]; *Liberty and others v the United Kingdom* App no 58243/00 (2008) 48 EHRR 1 [59]; *Yildirim v Turkey* App no 3111/10 (ECtHR, 18 March 2013) [57].
[69] *Yildirim v Turkey* App no 3111/10 (ECtHR, 18 March 2013) [59]; see also Concurring Opinion in *Yildirim v Turkey* App no 3111/10 (ECtHR, 18 March 2013), 27–28; *Barbulescu v Romania* App no 61496/08 (ECtHR, 5 September 2017) [133].
[70] 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]; Joined Cases C-203/15 and C-698/15 *Tele2 Sverige AB v Post-och telestyrelsenk* [2016] All ER (D) 107 (Dec) and *Secretary of State for the Home Department v Tom Watson* [2016] All ER (D) 107 (Dec) [121].
[71] 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.
[72] C-401/19 *Poland v Parliament and Council* [2022] ECLI:EU:C:2022:297 [24], [90].
[73] C-18/18 *Eva Glawischnig-Piesczek v Facebook Ireland Limited* [2019] ECLI:EU:C:2019:821 [41]-[46].

blocking, in safeguarding the right to freedom of expression, those platforms could not be required to undertake an independent assessment of the content, considering any copyright exceptions and limitations. It is true that Article 52(3)(2) explains that paragraph 3 does not apply if the use of deepfakes is allowed by law or if it is necessary to enjoy the right to freedom of expression and the right to freedom of the arts and sciences under Articles 11 and 13 EU Charter. If the content is artistic, satirical, fictional or creative filmmaking, video games visuals and the like, the transparency obligations are restricted to disclosing the deepfake in an adequate and visible clear way, not affecting the content display, and, if appropriate, disclosing the relevant copyrights. Problematically, however, as flagged above, the Act does not appear to specify any sanctions against those users failing to satisfy Article 52(3)(1) transparency obligations (EP 2021, 38).[74] Indeed, this can be contrasted with ECtHR case-law,[75] recognizing that the right to freedom of expression enshrined in Article 10 ECHR, is based on conditions, formalities, limitations, but perhaps more relevantly, also 'penalties as are prescribed by law'. Therefore, as the Act does not enable users to envisage the monitoring consequences of not labelling deepfake content nor contains any penalties for user non-adherence to Article 52(3)(1), it arguably fails to satisfy the ECtHR foreseeability principle under Articles 8(2) and 10(2).

In applying the rule of law principle, the ECtHR has observed that the applicability of the measures must also be subject to initial state authority oversight and appropriate safeguards.[76] As noted above, Recital 60h of the Act states that it is apt for the Commission and the AI Office to monitor and regularly evaluate the legal and governance framework of generative AI models that lead to significant questions regarding content creation, which infringes EU law, copyright regulations, and likely misuse. However, troublingly, disregarding ECtHR[77] and CJEU[78] case-law, the Act does not require the AI Office *prior* check and authorisation of specific generative AI systems, neither gives users effective safeguards against abuse. Arguably, before making specific generative AI models publicly available and deploying them, the AI Office should carefully review and licence them beforehand particularly when

---

[74] 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.

[75] *Yildirim v Turkey* App no 3111/10 (ECtHR, 18 March 2013) [38]; *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].

[76] *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]; *Rotaru v Romania* App no 28341/95 (2000) 8 BHRC 449 [59]; see also *Amann v Switzerland* App no 27798/95 (2000) 30 EHRR 843 [60].

[77] *Big Brother Watch and others v United Kingdom* App nos 58170/13, 62322/14 and 24960/15 (2018) ECHR 299 [318]; *Barbulescu v Romania* App no 61496/08 (ECtHR, 5 September 2017) [110]; *Klass and others v Germany* App no 5029/71 (1979–1980) 2 EHRR 214 [55]; *Rotaru v Romania* App no 28341/95 (2000) 8 BHRC 449 [59], [122]; see also *Amann v Switzerland* App no 27798/95 (2000) 30 EHRR 843 [60].

[78] Joined Cases C-203/15 and C-698/15 *Tele2 Sverige AB v Post-och telestyrelsenk* [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].

deepfake sextorsion, AI-child pornography or electoral disinformation content, is involved. In fact, in *Tele2/Watson* the CJEU found that, considering both the Court's case-law and Article 8(3) Charter, a key feature underpinning the protection of individuals regarding the processing of their personal data was *prior* review by the courts or independent authorities (ie AI Office).[79] Moreover, also conflicting with ECtHR[80] storage limitation and CJEU's[81] right to be forgotten case-law, *Clarkson v OpenAI* confirmed that yet nothing in the privacy policy of ChatGPT-4 demonstrated how the data, which had already been integrated into generative AI models, could ever truly be removed.[82] Therefore, the lack of notification of generative AI scraping methods, which detrimentally affected both adults and children (under thirteen), prevented users from exercising the right to request deletion. Potentially worse, OpenAI also failed to provide safeguards to ensure parental consent or that erasure of information concerning minors was ever possible.[83] Indeed, in *Kuric v Slovenia*, the ECtHR acknowledged how legislation which failed to clearly set out the significance of 'erasure', led to the claimants being unable to envisage the measure objected, but also to foresee its effect on their private life. It thus violated Article 8 ECHR.[84] In this context, arguably, the Act should expressly require providers tackling high-risk deepfakes created using generative AI to include in the privacy information: firstly, the purposes for data processing (e.g., detection of high-risk deepfakes); secondly, the retention periods for that personal and senstive data (e.g. up to 6 months); and lastly, any third-party sharing arrangements. As the UK's ICO stresses, if such providers gather data directly from individuals, they must give that privacy information to them when gathering it, but before using it to train the model or apply the model on them. Furthermore, if they gather it from additional sources, they must give that information within a reasonable time but not exceeding one month (ICO 2023, 28).[85] Thus, as overlooking ECtHR and CJEU case-law, the Act neither requires AI Office *prior* check and authorization of specific generative AI models nor affords adults and minors effective safeguards against abuse, arguably it fails to satisfy the ECtHR rule of law principle under Articles 8(2) and 10(2) ECHR.

---

[79] Joined Cases C-203/15 and C-698/15 *Tele2 Sverige AB v Post-och telestyrelsenk* [2016] All ER (D) 107 (Dec) and *Secretary of State for the Home Department v Tom Watson* [2016] All ER (D) 107 (Dec) [123].
[80] *S and Marper v the United Kingdom* App nos 30562/04 and 30566/04 (2008) ECHR 1581 [119], [124], [125]; *Gaughran v the United Kingdom* App no 45245/15 (ECtHR, 13 February 2020) [94]-[97]; *Brunet v France* App no 21010/10 (ECtHR, 18 Septemer 2014) [40]; *Aycaguer v France* App no 8806/12 (ECtHR, 22 June 2017) [42], [43].
[81] 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 Costeja González* [2014] ECLI:EU:C:2014:317.
[82] *PM et al v OpenAI LP*., 3:23-cv-03199 (US District Court, N.D. Cal. 2023) [311].
[83] *Ibid*. [274] and [490].
[84] *Kurić and Others v Slovenia* App no 26828/06 (ECtHR, 26 June 2012) [348], [349].
[85] ICO. 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/.

10.2. Legitimate aim

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.[86] The second prong 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, considering the ECtHR's findings in the ruling of *Glukhin v Russia*. It involved the Russian government's use of facial recognition.[87]

10.3. 'Necessary' and 'proportionate'

The next issue to be examined in this paper is to what extent under the Act, the provisions governing the use of deepfakes 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.[88] In addition, the Strasbourg Court has noted that the reasons provided by the state to justify the measures must be 'relevant and sufficient'.[89] 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.[90] This section will contend that the Act provisions governing the use of generative AI models facilitating the detection of deepfakes are unjustified, thereby infringing the Strasbourg Court's necessity and proportionality principles.

---

[86] *Delfi v Estonia* App no 64569/09 (ECtHR, 16 June 2015) [78]; *S and Marper v the United Kingdom* App no 30562/04 and 30566/04 (2008) ECHR 1581 [101]; *Coster v the United Kingdom* App no 24876/94 (2001) 33 EHRR 20 [104]; *Khurshid Mustafa and Tarzibachi v Sweden* App no 23883/06 (ECtHR, 16 March 2009) [43].

[87] *Glukhin v Russia* App no 11519/20 (ECtHR, 4 July 2023) [55], [84].

[88] *Delfi v Estonia* App no 64569/09 (ECtHR, 16 June 2015) [78]; *Cenzig and others v Turkey* App nos 48226/10 and 14027/11 (ECtHR, 1 December 2015) [58]; *Yildirim v Turkey* App no 3111/10 (ECtHR, 18 March 2013) [56]; *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].

[89] *Delfi v Estonia* App no 64569/09 (ECtHR, 16 June 2015) [78]; *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].

[90] *Delfi v Estonia App* no 64569/09 (ECtHR, 16 June 2015) [78]; *S and Marper v the United Kingdom* App no 30562/04 and 30566/04 (2008) ECHR 1581 [101]; *Coster v the United Kingdom* App no 24876/94 (2001) 33 EHRR 20 [104]; *Khurshid Mustafa and Tarzibachi v Sweden* App no 23883/06 (ECtHR, 16 March 2009) [43].

As far as the first principle is concerned, it is well-settled by Strasbourd case-law that, under Articles 8(2) and 10(2) ECHR, the measure's level of disruption is an important consideration when evaluating whether the methods used are necessary to achieving the desired outcome.[91] As outlined above, Recital 45a of the Act states that AI providers and users should adopt state-of-the-art organisational and technical measures including encryption and anonymisation, and use federated learning technology allowing multiple parties to train AI models on their own data without having to share the data with one another. However, in conflict with ECtHR[92] and CJEU[93] case-law, the Act fails spectacularly to illustrate how the deployment of high-risk AI, could be adopted in a less-privacy-and-data protection-invasive way. It is a fact that, Recital 45a stipulates that, under the GDPR, it is essential to follow the principles of data protection by design and by default, and data minimization when processing data that could pose significant risks to individual rights. Moreover, Article 58a elaborates that before using high-risk AI, deployers must also conduct both data protection impact assessments and fundamental rights impact assessments. Problematically, as previously highlighted, this however means that while deepfake detection deployers tackling electoral disinformation would have to evaluate the potential risks and harms showing why and how data processing is necessary and proportionate to achieve detection. Quite alarmingly, this would not be applied to sextorison or AI-generated child pornography situations. Confirming CJEU case-law on valid consent,[94] *Clarkson v OpenAI* emphasised that without any notification to the public, to train generative AI on which ChatGPT depended, no one could ever be considered to have consented to web-scraping, thereby being secretly monitored, profiled and targeted. Indeed, disturbingly, OpenAI created a powerful tool which was used to control user behaviour depriving everyone of meaningfully exercising their rights.[95] In this context, to comply with data subject GDPR rights,[96] it is suggested that providers of deepfake detection systems intended to tackle electoral disinformation, sextorison and AI-child pornography, should consider at the AI design stage, minimisation techniques such as structured synthetic data. This is a

---

[91] *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]; *Yildirim v Turkey* App no 3111/ 10 (ECtHR, 18 March 2013) [64]; *Uzun v Germany* App no 35623/05 (2010) 53 EHRR 852 [78].

[92] *Ibid*.

[93] 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]; C-83/14 CHEZ *Razpredelenie Bulgaria AD v Komisia za zashtita ot diskriminatsia* [2015] EUECJ [120]-[122].

[94] See for instance C-673/17 *Bundesverband der Verbraucherzentralen und Verbraucherverbände – Verbraucherzentrale Bundesverband e.V. v Planet49 GmbH* [2019] EU:C:2019:246 [61]; and C-61/19 *Orange Romania SA v Autoritatea Naţională de Supraveghere a Prelucrării Datelor cu Caracter Personal (ANSPDCP)* [2020] ECLI:EU:C:2020:901 [8], [39], [40], [52].

[95] *PM et al v OpenAI LP*., 3:23-cv-03199 (US District Court, N.D. Cal. 2023) [152], [248], [249], [269], [298].

[96] See Chapter 3 of the GDPR. Specifically, Article 12 i.e. transparent information, communication and modalities for the exercise of the rights of the data subject. Article 13 i.e. information to be provided where personal data are collected from the data subject. Article 14 i.e. information to be provided where personal data have not been obtained from the data subject. Article 15 i.e. right of access by the data subject. Article i.e. 16 right to rectification. Article 17 i.e. right to erasure ('right to be forgotten'). Article 18 i.e. right to restriction of processing. Article 19 i.e. notification obligation regarding rectification or erasure of personal data or restriction of processing. Article 20 i.e. right to data portability. Article 21 i.e. right to object. Article 22 i.e. automated individual decision-making, including profiling. Article 23 i.e. restrictions.

novel type of data employed for AI applications, enhancing the global AI training landscape, which addresses data minimization concerns, biased datasets or restricted data availability (Intel 2023).[97] Pursuant to the CJEU AG opinion in *Poland v Parliament and Council*, the deepfake detection error rate should however be as minimal as possible. Thus, if it were not feasible to use structured synthetic data without leading to a significant 'false positive' rate, such data filtering should be prohibited.[98] As noted above, Intel's real-time FakeCatcher can detect deepfake videos analysing blood flows with a 96 percent accuracy within milliseconds.[99] This may suggest that the false positive rate test is met. Thus, as the use of less-privacy-and-data protection-invasive structured synthetic data aimed at tackling deepfake electoral disinformation, sextorison and AI-child pornography, would have a less pronounced effect on provider and deployer rights, arguably the Act fails to satisfy the ECtHR necessity principle under Articles 8(2) and 10(2).

In applying the proportionality principle, the ECtHR has observed that legislation must also provide appropriate safeguards to afford individual protection against arbitrary interference.[100] As noted above, Article 28b of the Act states that generative AI providers shall design, train, and develop the model in a way, which guarantees appropriate safeguards against the creation of illegal content violating EU law, and without unduly restricting fundamental rights, such as the freedom of expression. However, it is worrisome that this questions ECtHR[101] and CJEU[102] case-law, echoing that to evaluate the necessity of a measure impacting privacy and data protection, it was also crucial to satisfy Article 5 of Convention 108, ensuring that data collection were relevant, accurate, adequate and not excessive, minimising the quantity of data gathered, ensuring transparency in their processing, and using them for specific and limited data collection purposes. One could thus argue that to create sufficient safeguards to prevent the creation of content that infringes EU law, the Act should also require deepfake providers to consider from the design stage different 'trade-offs'. Firstly, the interests in training a sufficiently accurate AI and, as analysed above, minimising the amount of personal data processed to train the model; secondly, creating AI that is sufficiently statistically accurate and which prevents

---

[97] 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.
[98] AG Opinion in C-401/19 *Poland v Parliament and Council* [2021] ECLI:EU:C:2021:613 [214].
[99] 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.
[100] *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]; *Rotaru v Romania* App no 28341/95 (2000) 8 BHRC 449 [59]; see also *Amann v Switzerland* App no 27798/95 (2000) 30 EHRR 843 [60].
[101] *Taylor-Sabori v the United Kigdom* 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]; *Mockutė v Lithuania* App no 66490/09 (ECtHR, 27 February 2018) [103]-[104]; *M.D. and Others v Spain* App no 36584/17 (ECtHR, 22 June 2022) [61]-[64].
[102] 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.

discrimination; and lastly, striking a fair balance between statistically accuracy, explainability, commercial secrecy, and security (ICO 2023, 24).[103] It is undeniable that, based on Recital 71 GDPR, while 'statistical accuracy' concerns the AI accuracy itself, deepfake detection systems need not be 100 percent statistically accurate to observe the accuracy principle. Arguably, however, recognising the CJEU AG warning in *SCHUFA (Scoring)*, deepfake detection providers should also be obliged to provide meaningful information regarding the system's logic, including sufficiently comprehensive explanations of the techniques deployed to measure the detection score and the rationale behind a specific outcome. As discussed above, when tackling electoral disinformation, AI-child pornography, and sextorsion, this could potentially entail the reasons for a significant or insignificant 'false positive' rate result. In *SCHUFA (Scoring)* the AG concluded that, individuals should also be given general information, on the factors considered for the decision-making process and their corresponding weight in total, to contest any decision, acknowledging the GDPR right not to be subject to a decision based solely on automated processing, including profiling.[104] Moreover, *Clarkson v OpenAI* exposed how despite 'Open' AI's 'absolute secrecy' regarding its data gathering and practices, generative AI could also target vulnerable people with algorithmic discrimination, and predatory advertising, also maximising human bias due to training its generative AI models with harmful content.[105] Indeed, in addition to creating undetectable malware to compromise security systems, OpenAI also threatened both domestic and international security as 'killer robots', could detect, select, and assassinate humans without individual intervention.[106] Thus, as the Act neither requires deepfake detection providers to implement sufficiently statistically accurate systems, which avoid discrimination, nor a fair balance to be struck between explainability, statistically accuracy, security, and commercial secrecy, it arguably fails to satisfy the ECtHR proportionality principle under Articles 8(2) and 10(2).

Regarding the proportionality principle, the ECtHR has explained that monitoring and technical measures must also strike a fair balance between the competing interests at stake.[107] As noted above, Article 28b of the Act elaborates that generative AI providers and those implementing foundation models into generative

---

[103] ICO. 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/.
[104] AG Opinion in C-634/21 *SCHUFA Holding and Others (Scoring)* [2023] ECLI:EU:C:2023:220 [AG 58].
[105] *PM et al v OpenAI LP*., 3:23-cv-03199 (US District Court, N.D. Cal. 2023) [151], [202], [228].
[106] *Ibid.* [229], [236].
[107] *Glukhin v Russia* App no 11519/20 (ECtHR, 4 July 2023) [56], [57]; *Barbulescu v Romania* App no 61496/08 (ECtHR, 5 September 2017) [140]–[141]; Concurring Opinion in *Yildirim v Turkey* App no 3111/10 (ECtHR, 18 March 2013) page 29; *Delfi v Estonia* App no 64569/09 (ECtHR, 16 June 2015) [159].

AI, shall additionally document and make publicly available a sufficiently comprehensive report on the use of training data, which is protected under domestic or EU copyright law. Controversially, however, contradicting ECtHR[108] and CJEU[109] case-law, the Act appears to disregard the fact that, in addition to copyright-related matters, generative AI providers should also be obliged to document and publicly report on training data use, which affects not only users' privacy, data protection and intellectual property rights, under Articles 7, 8 and 17(2) Charter, but also AI companies' trade secret protection, and freedom to conduct their business, under Articles 17(2), and 16 Charter. Firstly, OpenAI's technical report reveals how ChatGPT-4 was trained on publicly available information like online data and third-party provider licensed one (OpenAI 2023, 2).[110] Undeniably, in *Poland v Parliament and Council* and *UPC Telekabel Wien* the CJEU found that, to strike a fair balance, it was essential under Article 16 Charter, to leave providers to decide the specific measures to be adopted considering their available abilities and resources.[111] Notably, however, *Clarkson v OpenAI* also uncovered how the commercial misappropriation of the 'Common Crawl' database, upon which OpenAI depended, originated from massive data extraction.[112] This involved millions of users' personal information without their knowledge or consent and violated both their property and privacy rights, constituting theft.[113] Moreover, to train ChatGPT, OpenAI also deployed training datasets, which comprised any information a user prompted, along with that user's contact details, account information, IP addresses, login credentials, and other sensitive data including cookies and analytics.[114] Indeed, *Clarkson v OpenAI* also warned that by enabling the gathering, retention, and analysis of a tremendous amount of individualised, personal and sensitive data, from audio and visual data to specific preferences, habits and interests; frighteningly, generative AI rapidly accelerated the widespread of 'deepfakes'.[115] Recital 60f of the Act offers generative AI deepfake detection providers the option between granting direct access to users to their model outputs, but also its training methods and algorithms, or else: facing 'appropriate risk mitigation' for any

---

[108] *Delfi v Estonia* App no 64569/09 (ECtHR, 16 June 2015) [13]; *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]; *Yildirim v Turkey* App no 3111/10 (ECtHR, 18 March 2013) [28], [64].

[109] C-401/19 *Poland v Parliament and Council* [2022] ECLI:EU:C:2022:297 [75], [99]; C-18/18 *Eva Glawischnig-Piesczek v Facebook Ireland Limited* [2019] ECLI:EU:C:2019:821 [43];C-314/12 *UPC Telekabel Wien GmbH v Constantin FilmVerleih GmbH and Wega Filmproduktionsgesellschaft GmbH* [2013] EU:C:2014:192 [52]; 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]; C-275/06 *Productores de Musica de Espana (Promusicae) v Telefonica de Espana SAU* [2008] ECR I-271 [68].

[110] OpenAI. 2023. "GPT-4 technical report." https://cdn.openai.com/papers/gpt-4.pdf.

[111] C-401/19 *Poland v Parliament and Council* [2022] ECLI:EU:C:2022:297 [75]; C-314/12 *UPC Telekabel Wien GmbH v Constantin FilmVerleih GmbH and Wega Filmproduktionsgesellschaft GmbH* [2013] EU:C:2014:192 [52].

[112] *PM et al v OpenAI LP*., 3:23-cv-03199 (US District Court, N.D. Cal. 2023) [155].

[113] *Ibid.* [247], [258].

[114] *Ibid.* [151], [163].

[115] *Ibid.* [219].

'downstream provider' to which it provides access. Arguably, however, the first option would be unworkable for most providers as their whole business model frequently depends on developing proprietary AI. The second alternative would also require providers to monitor everyone deploying their services and implement adequate safeguards (Wolff et al. 2023, 8-9).[116] Unfortunately, this would not only conflict with CJEU's case-law protecting AI companies' commercial secrecy,[117] but also prohibiting them from undertaking general user monitoring obligations.[118] Thus, as the Act fails to strike a fair balance between users' privacy, data protection and intellectual property rights, and AI providers' trade secret protection, and freedom to conduct their business, a case can be made that it fails to satisfy the ECtHR proportionality principle under Articles 8(2) and 10(2).

11. Discussion of findings

In the revolutionary era of artificial intelligence, the potential impact of generative AI on human rights is a major research area among legal scholars. A number of studies have examined the possible benefits of using human rights as a benchmark for evaluating such impact (Sison et al. 2023,[119] Wolff et al. 2023,[120] Helberger and Diakopoulos. 2023,[121] Human Rights Watch. 2023,[122] Lucchi. 2023[123]). Surprisingly, however, there has been limited research on the compatibility of the EU AI Act provisions governing deepfakes with Articles 8 and 10 ECHR, and the GDPR. This paper makes a significant contribution to the field. It suggests that the Act be amended to include new provisions, requiring AI providers to use structured synthetic data to detect deepfakes, and along with electoral disinformation, also explicitly add to the list of high-risk AI, systems used for AI-child pornography and sextorsion. The findings of this research are supported by the case-law of the Strasbourg and Luxembourg courts. In *Glukhin v Russia* the ECtHR emphasised

---

[116] Wolff, T., William Lehr, Christopher Yoo. 2023. "Lessons from GDPR for AI policy making." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4528698.

[117] C-54/21 *Antea Polska S.A., Pectore-Eco sp. z o.o., Instytut Ochrony Środowiska — Państwowy Instytut Badawczy v Państwowe Gospodarstwo Wodne Wody Polskie* [2022] ECLI:EU:C:2022:888 [51]-[55]; C-927/19 *Klaipėdos regiono atliekų tvarkymo centras v UAB* [2021] ECLI:EU:C:2021:700 [49], [50], [62], [63].

[118] C-401/19 *Poland v Parliament and Council* [2022] ECLI:EU:C:2022:297 [86], [90]; C-18/18 *Eva Glawischnig-Piesczek v Facebook Ireland Limited* [2019] ECLI:EU:C:2019:821 [46], [53]; C-484/14 *Tobias Mc Fadden v Sony Music Entertainment Germany GmbH* [2016] ECLI:EU:C:2016:689 [87]; C-360/10 *Belgische Vereniging van Auteurs, Componisten en Uitgevers CVBA (SABAM) v Netlog NV* [2012] ECLI:EU:C:2012:85 [33]-[38]; C-70/10 *Scarlet Extended SA v Société belge des auteurs, compositeurs et éditeurs SCRL (SABAM)* [2012] ECLI:EU:C:2011:771 [35]-[40]; C-324/09 *L'Oréal SA and others v eBay International AG and others* [2011] ECLI:EU:C:2011:474 [139]-[140].

[119] Sison, A.J., Marco Tulio Daza, Roberto Gozalo-Brizuela, Eduardo César Garrido Merchán. 2023. "ChatGPT: more than a "weapon of mass deception" ethical challenges and responses from the human-centered artificial intelligence (HCAI) Perspective." *International Journal of Human-Computer Interaction*. https://www.tandfonline.com/doi/abs/10.1080/10447318.2023.2225931.

[120] Wolff, T., William Lehr, Christopher Yoo. 2023. "Lessons from GDPR for AI policy making." https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4528698.

[121] Helberger, N. and Diakopoulos, N. 2023. "ChatGPT and the AI Act." *Internet Policy Review*. https://policyreview.info/essay/chatgpt-and-ai-act.

[122] Human Rights Watch. 2023. "Pandora's box: generative AI, companies, ChatGPT, and human rights." https://www.hrw.org/news/2023/05/03/pandoras-box-generative-ai-companies-chatgpt-and-human-rights.

[123] Lucchi, N. 2023. "ChatGPT: a case study on copyright challenges for generative artificial intelligence systems." *Cambridge University Press*. https://www.cambridge.org/core/journals/european-journal-of-risk-regulation/article/chatgpt-a-case-study-on-copyright-challenges-for-generative-artificial-intelligence-systems/CEDCE34DED599CC4EB201289BB161965.

that, national legislation on personal data gathering and processing must have detailed, clear norms specifying the application and scope of measures, along with minimum safeguards.[124] The ECtHR strongly doubted whether the domestic framework governing the use of AI-based facial recognition satisfied the 'quality of law' condition.[125] However, it highlighted that it was exclusively required to establish whether the applicant's data processing was 'necessary'.[126] It found that the deployment of facial recognition to detect Glukhin violated Article 8 ECHR.[127] Similarly, observing *Promusicae*,[128] in *Poland v Parliament and Council,* the CJEU confirmed that while specifically targeted filtering and blocking measures, which adequately distinguished between lawful and unlawful content might be deployed, general monitoring obligations imposed on providers to take preventive measures against future infringements were indeed prohibited.[129] Moreover, when implementing such measures, domestic legislation needed to allow a fair balance to be struck between all the relevant interests, but also ensure that they respected the Charter rights, including the principle of proportionality.[130]

Notably, the paper's findings could also have implications for society, the economy and the environment. However, the decision as to whether demanding deepfake detection providers to utilise structured synthetic data, and the Act narrowly target electoral disinformation, AI-child pornography, and sextorion content, must also be justified considering such impact. Recital 44 of the Act explains how AI-generated outomes can be susceptible to inherent biases which tend to gradually widen the gap between different groups, thereby reinforcing and strengthening existing discrimination, particularly for individuals belonging to specific ethnic or vulnerable communities, or radicalised groups. Indeed, AI-based detection systems might not always be effective in identifying individuals with darker skin tones (Google AI Skin Tone 2022),[131] and yet if the training datasets do not include all skin tones, ethnicities, genders, ages and accents, such systems could also lead to false positives. For instance, research indicates that AI identification tools tend to be biased towards white middle-aged men, but worryingly; also negatively impacting on marginalised groups (The Guardian News, August 17, 2023). In this regard, by reducing discriminatory results, structured synthetic data would enable providers of

---

[124] *Glukhin v Russia* App no 11519/20 (ECtHR, 4 July 2023) [77].
[125] *Ibid*. [83].
[126] *Ibid*. [85], [86].
[127] *Ibid*. [89], [90], [91]; in this context, concerning the deployment of AI-powered facial recognition technology such as, Clearview AI and its impact on human rights see also Romero-Moreno, F. 2022. ''Facial recognition technology: how it's being used in Ukraine and why it's still so controversial.'' *The Conversation*. https://theconversation.com/facial-recognition-technology-how-its-being-used-in-ukraine-and-why-its-still-so-controversial-183171.
[128] C-275/06 *Productores de Musica de Espana (Promusicae) v Telefonica de Espana SAU* [2008] ECR I-271 [68].
[129] C-401/19 *Poland v Parliament and Council* [2022] ECLI:EU:C:2022:297 [86], [90].
[130] *Ibid.* [99].
[131] Google AI Skin Tone Research. 2022. ''Skin Tone Research @ Google AI.'' https://skintone.google/.

deepfake detection systems to create large quantities of diverse data and generate representative or balanced samples, which more accurately represent the target population. Importantly, such data would also solve the overfitting problem, where the detection system works well on the training data, but it may not be able to classify new, unseen deepfake content. Moreover, classic data gathering techniques are resource-demanding, expensive, and time-consuming. Therefore, the use of structured synthetic data, would also decrease the costs related to data gathering and retention, thus enabling providers to concentrate on the analysis instead of data collection. Additionally, due to its authentication and privacy-preserving features, this data would also protect data security allowing providers to create a synthetic dataset that mimics real-world data without revealing any sensitive information (Syntheticus 2023).[132]

Furthermore, regarding environmental impact, research discloses that the training of 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).[133] However, problematically, neither OpenAI's technical report,[134] nor *Clarkson v OpenAI*,[135] unmasked the carbon footprint of developing ChatGPT-4. Yet, while one large-scale generative AI model would not cause much environmental damage, if numerous companies were to develop models with slight differences for various purposes such as detecting sextorison, electoral disinformation and AI-child pornography, each used by millions of users; undeniably the high demand for energy could become unsustainable (Saenko. 2023).[136] Article 87a of the Act correctly notes that there is limited and reliable information on energy use, including hardware, software and especially datacenters. Thus, it proposes that the Commission should adopt an appropriate methodology to assess the environmental impact of AI systems. Arguably, however, the undertaking of any environmental impact assessment to evaluate CO2e may be difficult to achieve as it is hard to quantify all the needed information because it is not readily available or accessible including hardware, datacenters and energy mix, but also challenging to disclose essential information subsequently (Patterson et al. 2021,

---

[132] 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.

[133] Patterson, D., Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-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.

[134] OpenAI. 2023. "GPT-4 technical report." https://cdn.openai.com/papers/gpt-4.pdf.

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

[136] 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.

2).[137] In this context, it is suggested that, when developing the guidelines which consider the environmental impact of generative AI, including energy efficiency and carbon footprint under Article 82b(h), the Commission and the AI Office, should support, and adopt Google's leading approach in this area. Indeed, Google's business model has long prioritised improving the energy efficiency of algorithms, software, hardware, and datacenters. For example, Google Cloud allows clients to choose the datacenter based on CO2e, and publishes updates on the level of carbon-free energy and gross CO2e of these datacenters (Google Cloud 2023).[138]

Moreover, Recital 60h elaborates that generative AI internal assessments should also consider industry standards and concentrate on acquiring adequate technical knowledge and control over the model. Therefore, in addition to acknowledging the paper recommendations, arguably deepfake detection providers should also adopt a comprehensive approach to addressing harmful content. Research recommends that to prevent the widespread effect of deepfakes, one solution is to integrate detection methods into social media platforms using monitoring, filtering and blocking (Thi Nguyen et al. 2022, 13).[139] Indeed, Recital 12 clarifies that the Act should apply without impacting the intermediary provider liability regime, under the EU DSA. This indicates that to tackle deepfakes, such platforms could also adopt notice-and-takedown and notice-and-staydown systems (Romero-Moreno 2019),[140] and (Romero-Moreno 2020).[141] Moreover, it is also possible to create AI-generated content including watermarks, which identify it as synthetic. By incorporating the watermark into generative AI during the training process, the AI-generated content will also include the same watermark. The ideal watermark should however be invisible and able to withstand removal or modification attempts, including colour adjustment, resizing, cropping and compression (Farid. 2023).[142] In this regard, it is eye-catching Google Deepmind's SynthID, which integrates watermarks into AI-generated images by embedding changes to pixels, notably even if one were to alter the image's colour, contrast, or dimensions (Google DeepMind 2023).[143] Yet,

---

[137] Patterson, D., Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-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.
[138] Google Cloud. 2023. ''Carbon free energy for Google Cloud regions.'' https://cloud.google.com/sustainability/region-carbon; see also *ibid* at page 15.
[139] Thi Nguyena, T., Quoc Viet Hung Nguyenb, Dung Tien Nguyena, Duc Thanh Nguyena, Thien Huynh-Thec, Saeid Nahavandid, Thanh Tam Nguyene, Quoc-Viet Phamf, Cuong M. Nguyeng. 2022. "Deeplearning for deepfakes creation and detection: a survey." *Computer Vision and Image Understanding*. https://www.sciencedirect.com/science/article/abs/pii/S1077314222001114.
[140] 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/10.1080/13600869.2018.1475906.
[141] Romero-Moreno, F. 2020. "'Upload filters' and human rights: implementing Article 17 of the EU Copyright Directive in the Digital Single Market.'' *International Review of Law, Computers and Technology*. https://www.tandfonline.com/doi/full/10.1080/13600869.2020.1733760.
[142] Farid, H. 2023. "Watermarking ChatGPT, DALL-E and other generative AIs could help protect against fraud and misinformation." *The Conversation*. https://theconversation.com/watermarking-chatgpt-dall-e-and-other-generative-ais-could-help-protect-against-fraud-and-misinformation-202293.
[143] Google DeepMind. 2023. "Identifying AI-generated images with SynthID." https://www.deepmind.com/blog/identifying-ai-generated-images-with-synthid

although such technical solutions can be a deterrent to abuse, studies reveal how to easily bypass them (Ghandi and Jain. 2020, 1-8).[144] Lastly, research concludes that while using detection systems is important, it is perhaps even more critical to understand the reasons why individuals publish deepfakes including who shared it and what was the public's reaction (Intelligencer. 2019).[145] However, despite being trained, humans are still only able to identify deepfake speech with 70% accuracy (The Guardian News, August 2, 2023). In sum, the fight against deepfakes is a long-term arms race, the techniques used to create, share and detect deepfakes are constantly being improved. Therefore, arguably one may need to do the same.

12. Conclusion

This paper has critically assessed the extent to which under the EU AI Act the provisions governing the use of deepfakes could be adopted in a way that is consistent with the right of AI providers and users to privacy and freedom of expression under Articles 8 and 10 of the Convention, and the GDPR. The paper addresses a significant gap in the literature. It proposes that the Act be amended to introduce new obligations for AI providers oblige them to deploy structured synthetic data to detect deepfakes, and in addition to electoral disinformation, also explicitly consider AI systems intended to be used for sextorsion and AI-child pornography, high-risk AI. I conclude that unless, pursuant to Article 7(1), empowering the Commission to amend the Act, the procedural safeguards recommended below are implemented through delegated acts, its provisions reguating deepfakes will violate Articles 8 and 10 ECHR, and the GDPR.

- In conflict with the ECtHR principle of accessibility, apart from electoral disinformation, the Act fails to consider, much less recognise other specific types of deepfake high-risk AI systems, thus AI providers and users being unable to clearly and easily understand how these systems will affect them. The first procedural safeguard should therefore be for the Act to explicitly add to the list of high-risk AI, deepfake sextorsion and AI-child pornography, thereby providing effective means to protect their rights.

- In the way it ignores the ECtHR principle of foreseeability, the Act offers no guidance on whether the platform, should play a role in facilitating and/or

[144] Gandhi, A., and Jain. S. 2020. "Adversarial perturbations fool deepfake detectors." *ArXiv*. https://arxiv.org/pdf/2003.10596.pdf.
[145] 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.

monitoring the labelling of deepfake content. Neither Article 71 spells out among the penalties whether user non-adherence to Article 52(3)(1) transparency obligations, is sanctionable (EP 2021, 38).[146] The second procedural safeguard to be adopted is that the Act should expressly address the implications of not labeling deepfakes and the penalties for users who do not comply with Article 52(3)(1).

- By overlooking the ECtHR principle of rule of law, the Act does not require AI Office *prior* check and authorization of specific generative AI models. It also does not provide adults and minors effective safeguards against abuse. The third procedural safeguard should thus be for the use of generative AI involving sextorsion, AI-child pornography and electoral disinformation to be subject to AI Office initial check and authorization, and ensure users appropriate safeguards such as, notification, right to deletion and parental consent.

- Contrary to the ECtHR principle of necessity, the Act fails spectacularly to illustrate how the use of high-risk AI systems, could be adopted in a less intrusive way to AI provider and user privacy and data protection rights. A further procedural safeguard to be put in place is that the Act should require AI providers to deploy less-privacy-and-data protection-invasive structured synthetic data specifically targeted at deepfake electoral disinformation, sextorison and AI-child pornography content.

- In conflict with the ECtHR principle of proportionality, the Act does not meet the requirements set out in Article 5 of Convention 108. An additional procedural safeguard should therefore be for deepfake detection providers to use sufficiently statistically accurate and unbiased systems, which strike a fair balance between explainability, statistically accuracy, security, and commercial secrecy.

[146] 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.

- By disregarding the ECtHR principle of proportionality, under the Act generative AI providers are only obliged to document and make publicly available sufficiently comprehensive reports on training data usage, considering copyright law. The last procedural safeguard to be adopted is that the Act should afford all the interests involved equal weight and consideration including users' privacy, data protection and IP rights, and AI providers' trade secret protection, and freedom to conduct their business.

Article 7(1) is the provision in the Act that has been given the utmost care and consideration. In fact, one could argue that it has been painstakingly crafted. It enables the Commission to implement delegated legislation to amend Annex III by adding or changing areas or applications of high-risk AI, if specific systems pose a significant risk of harm to fundamental rights like deepfake sextorison, AI-child pornography, and electoral disinformation. Thus, the Commission would be well-advised to adopt these recommended safeguards as they would ensure the Act is implemented in a way that protects the rights of providers and users.

In my view, however, if the suggested procedural safeguards are not taken into account during the consultation process for amendments in the preparation of delegated acts pursuant to Recital 85, no other position would be more likely to ignite anger and resentment among consumer groups, civil society, organizations representing affected individuals, executives from small, medium, and large businesses, as well as scientists and researchers. Indeed, alarmingly, the AI Act deepfake provisions would be violating the right of AI providers and users to privacy and freedom of expression under Articles 8 and 10 of the Convention, and the GDPR.