"Deep fakes": disentangling terms in the proposed EU Artificial Intelligence Act

Angelica Fernandez[*]

I. Introduction 395
1. The state of deep fake phenomena 395
2. Overview of current regulatory context 398
3. Summary of the proposed AI Act 400
II. Defining deep fakes: issues and challenges 402
1. Building consensus for a legal definition of deep fakes 402
a) Scholarly perspectives ... 402
b) Industry stakeholders' perspectives 404
c) Elements of consensus ... 405
2. Challenges to a practical definition 406
a) Setting the boundary between deep fakes and other forms of audiovisual manipulation 406
b) The "Liar's dividends" effects of deep fakes 407
3. Further elements of analysis 408
a) A technology driven approach to deep fake regulation 408
b) A distinct profile harm for deep fakes 410
III. An approach to regulating deep fakes in the proposed AI Act 411
1. Transparency requirements for deep fakes 412
2. Labelling requirements for deep fakes 413
3. Lessons from the Code of Practice on Disinformation 414
4. Other normative concerns regarding deep fakes 417
a) Normative concerns over enforcement mechanisms 417
b) Concerns over dual use of deep fakes in a risk-based assessment framework 418
IV. Criminalizing malicious deep fakes 420
1. Deep fakes and revenge porn legislations 420
2. Legal responses to deep fake phenomena: selected examples of non-EU jurisdictions 423
a) United States 423
b) United Kingdom 426
V. Conclusion 427

* *Angelica Fernandez* is a PhD candidate at the Department of Law of the Faculty of Law, Economics and Finance of the University of Luxembourg supported by the Luxembourg National Research Fund (FNR) (PRIDE17/12251371). Contact at: angelica.fernandez@uni.lu.

UFITA 2/2021, DOI: 10.5771/2568-9185-2021-2-392

Abstract

Since 2018, deep fakes technology has been one of the areas in which artificial intelligence has evolved dramatically and thus, deep fakes are primarily seen by governments as an emerging threat. In particular, regulators are increasingly concerned by the developments and applications of this technology in two main areas: image-based sexual abuse and disinformation. Despite its increasing popularity, there are challenges in defining what deep fakes are and what ought to be regulated when it comes to deep fake phenomena.

The following article aims to analyze the EU-level regulatory approach to deep fakes in relation to AI regulation. This choice is motivated by the inclusion of deep fakes in the proposed EU Artificial Intelligence Act and the nature of the provisions that apply to deep fake technology within the Act. The first part will analyze the issues and challenges of adopting a legal definition for deep fakes to highlight consensus and differences among scholars and industry players. Getting the scope of the definition right is essential to address appropriately the distinct harm profile stemming from deep fake technology, specifically in relation to image-based sexual abuse and disinformation. A survey of different views shows a consensus over two elements that define deep fakes: the use of AI-based technology and the intent of the creator. However, there are practical challenges to this seemingly consensual definition, particularly when it comes to drawing boundaries between deep fakes and lower AV manipulation (i.e.: cheap fakes) and to co-opting the term to discredit audiovisual content, and casting doubts on the veracity of AV content presented as evidence.

The second part focuses on the transparency requirements for deep fakes under the proposed EU Artificial Intelligence Act proposal. This obligation will be examined in light of disclosure and labelling obligations already tested in disinformation strategies, particularly in the implementation of the EU Code of Practice on Disinformation 2018, which will likely include deep fakes in its new iteration to be published in spring 2022. Among the main lessons from the application of the Code of Practice on Disinformation, it is clear that labels alone are not an effective measure to counter disinformation or deter its creation and dissemination. Moreover, if users are to rely on labels to weigh whether they are interacting with manipulated media, more research is needed into effective design since newer forms of enhancing transparency are available but not necessarily implemented by companies. This is particularly relevant in the context of the proposed Artificial Intelligence Act, since scholars have serious concerns of its enforcement architecture.

Finally, the third part of this article illustrates a brief comparative approach between the United States and the United Kingdom regulatory responses to deep fakes to assess further the current EU response to this phenomenon. In contrast to the EU response, which so far is based on minimal transparency requirements, other jurisdictions' trend has been primarily to criminalize the malicious use of deep fakes which is often assimilated to revenge pornography even though these are two different phenomena. For all of these reasons, deep fakes are at the intersection of different possible regulatory frameworks providing an interesting case to explore the regulatory challenges of AI in the context of the European Union.

Seit 2018 ist die Deep-Fake-Technologie einer der Bereiche, in denen sich die künstliche Intelligenz (KI) dramatisch weiterentwickelt hat. Aus diesem Grund werden Deep-Fakes von Regierungen in erster Linie als eine aufkommende Bedrohung angesehen. Regulierungsbehörden sind insbesondere zunehmend besorgt über die Entwicklungen und Anwendungen dieser Technologie in zwei Hauptbereichen: bildbasierter sexueller Missbrauch und Desinformation. Trotz der zunehmenden Beliebtheit dieser Technologie ist es schwierig zu definieren, was Deep Fakes sind und was reguliert werden sollte, wenn es um Deep-Fake-Phänomene geht.

Der folgende Artikel zielt darauf ab, den regulatorischen Ansatz zu Deep Fakes auf EU-Ebene in Bezug auf die Regulierung von KI zu analysieren. Der Grund für diese Entscheidung ist die Aufnahme von Deep Fakes in das vorgeschlagene EU-Gesetz über Künstliche Intelligenz und die Beschaffenheit der in diesem enthaltenen, in Bezug auf Deep-Fake-Technologien geltenden Bestimmungen. Im ersten Teil werden die Probleme und Herausforderungen bei der Festlegung einer rechtlichen Definition für Deep Fakes analysiert, um Konsens und Unterschiede zwischen Wissenschaftlern und Branchenvertretern herauszustellen. Der richtige Definitionsumfang ist von entscheidender Bedeutung, um dem ausgeprägten Schadenspotenzial der Deep-Fake-Technologie gerecht zu werden, insbesondere in Bezug auf bildbasierten sexuellen Missbrauch und Desinformation. Eine Übersicht über die verschiedenen Meinungen zeigt einen Konsens über zwei Elemente, die Deep Fakes definieren: die Verwendung von KI-basierter Technologie und die Absicht des Urhebers. Diese scheinbar einvernehmliche Definition birgt jedoch praktische Probleme, vor allem wenn es darum geht, die Grenzen zwischen Deep Fakes und niedrigschwelligeren audiovisuellen Manipulationen (d.h. billigen Fälschungen) zu ziehen und den Begriff zu vereinnahmen, um audiovisuelle Inhalte zu diskreditieren und den Wahrheitsgehalt audiovisueller Inhalte, die als Beweise vorgelegt werden, in Zweifel zu ziehen.

Der zweite Teil konzentriert sich auf die Transparenzanforderungen für Deep Fakes im Rahmen des vorgeschlagenen EU-Gesetz über Künstliche Intelligenz. Diese Verpflichtung wird im Lichte der Offenlegungs- und Kennzeichnungspflichten untersucht, die bereits im Rahmen von Desinformationsstrategien erprobt wurden, insbesondere bei der Umsetzung des EU-Verhaltenskodex zur Bekämpfung von Desinformation 2018, der in seiner neuen Fassung, die im Frühjahr 2022 veröffentlicht werden soll, voraussichtlich auch Deep Fakes umfassen wird. Eine der wichtigsten Lehren aus der Anwendung des Verhaltenskodex zur Bekämpfung von Desinformation ist, dass Kennzeichnungen allein keine wirksame Maßnahme sind, um Desinformation zu bekämpfen oder ihre Erstellung und Verbreitung zu verhindern. Wenn sich die Nutzer auf Kennzeichnungen verlassen sollen, um abzuwägen, ob sie mit manipulierten Medien interagieren, ist außerdem mehr Forschung über eine wirksame Ausgestaltung erforderlich, da neuere Formen zur Verbesserung der Transparenz zwar verfügbar sind, aber nicht unbedingt von den Unternehmen umgesetzt werden. Dies ist insbesondere im Zusammenhang mit dem vorgeschlagenen Gesetz über Künstliche Intelligenz

von Bedeutung, da Wissenschaftler ernsthafte Bedenken hinsichtlich seiner Durchsetzungsarchitektur haben.
Schließlich wird im dritten Teil dieses Artikels ein kurzer Vergleich zwischen den regulatorischen Reaktionen der Vereinigten Staaten und des Vereinigten Königreichs auf Deep Fakes angestellt, um die derzeitige Reaktion der EU auf dieses Phänomen zu bewerten. Im Gegensatz zur Reaktion der EU, die sich bisher auf minimale Transparenzanforderungen stützt, geht der Trend in anderen Ländern vor allem dahin, die böswillige Nutzung von Deep Fakes zu kriminalisieren, die oft mit Rache-Pornografie in Verbindung gebracht wird, obwohl es sich um zwei unterschiedliche Phänomene handelt. Aus all diesen Gründen befinden sich Deep Fakes im Schnittpunkt verschiedener möglicher regulatorischer Rahmenbedingungen und stellen einen interessanten Fall dar, um die regulatorischen Herausforderungen der KI im Kontext der Europäischen Union zu untersuchen.

I. Introduction

1. The state of deep fake phenomena

In 2019 the insurance firm *Euler Hermes Group SA* told the story of an anonymous energy company based in the UK that fell prey to an audio deep fake and ended up transferring €220,000 to the perpetrators of the crime.[1] Something to worry about for the insurance company facing these new types of cybercrimes cases and recovering the company losses. In this case, the company CEO in the UK thought he was speaking with the chief executive of the parent company based in Germany, who ordered him to transfer the funds. The authenticity of the spoof audio was such that the unknown number raised suspicion rather than the voice itself, confirming the high-quality one could obtain with commercial voice-generating software. The previously poor reputation of synthetic audio used for cybercrime or scam calls is increasingly changing. After 2020, due to the COVID-19 pandemic, online interactions have become more ubiquitous, and more companies are looking to brand themselves through a distinct voice, as they have done in the past with visual logos to create a unique visual commercial identity. Many start-ups offer generating voices for corporate e-learning videos, digital assistants, call center operators, video-game characters, and personalized advertising.[2]

The increasing popularity of synthetic voice is matched in the area of synthetic video production. Media outlets have reported on high-profile cases such as the 2018 "Obama PSA" lip-synch deep fake, in which actor *Jordan Peele* ventriloquizes *Barack Obama* to

1 *BBC News*, 'Fake Voices "Help Cyber-Crooks Steal Cash"' (8 July 2019) <https://www.bbc.co.uk/news/technology-48908736> accessed 1 December 2021. All URLS for this article were last accessed on 1st December 2021.

2 *Karen Hao*, 'AI Voice Actors Sound More Human than Ever—and They're Ready to Hire' (2019) MIT Technology Review <https://www.technologyreview.com/2021/07/09/1028140/ai-voice-actors-sound-human/>.

having him voice his opinion. This demonstration raised for the first time a mass awareness on the potential of deep fake technology. Currently, the *Montreal Institute for Learning Algorithms* (MILA) has developed a first model of what they call "ObamaNet",[3] which is a system that can generate videos of a person reading aloud any arbitrary text given to the system. The result is natural and realistic deep fakes lip-synching videos. Another prominent field of development for commercial uses of deep fakes has been the entertainment industry. Traditionally, computer-generated imagery (CGI) technology was considered the standard in this industry to create realistic images. However, famously, a fan showed that deep fake technology can surpass movie studio CGI when he published a reenactment video of princess Lea in the "Rogue One" *Star-Wars* movie, which was considered more realistic than the original version which used CGI. [4] More problematic, the media has also reported over the last years on the many celebrities' fake non-consensual pornographic videos or the different apps available to create them.[5] This variety of examples illustrates the dual-use nature of deep fake technology, which is a major challenge when thinking about regulating the uses of this technology.

Since 2018, according to a report on the state of deep fakes the "use of AI to generate harmful synthetic video, images, or audio, popularized under the broadername of 'deepfakes'" is one of the areas in which AI has evolved dramatically.[6] *Sensity*, a European deep fake detection company, estimates that in 2019 14,678 deep fakes were circulating on the Internet, which was almost a 50% increase from 2018 when deep fakes were first acknowledged, and almost the triple from 2017 when the term was first coined on a *Reddit* forum.[7]

3 See *Rithesh Kumar*, 'ObamaNet: Photo-Realistic Lip-Sync from Text' (NeurIPS, 2017) <https://ritheshkumar.com/obamanet/>.

4 Computer-generated imagery technology was until recently the standard technology used by movie studios to produce high-quality audiovisual manipulation. It was first used for the *Star Wars* movie saga and later developed to become mainstream to all movie studios. The movie *Avatar* directed by *James Cameron* in 2009 was one of the most complex early undertakings of CGI. But since 2006 most *Disney* and *Marvel Studios* films, for example, use CGI in variable degrees.

5 *Joseph Foley*, '14 Deepfake Examples That Terrified and Amused the Internet' (Creative Bloq) <https://www.creativebloq.com/features/deepfake-examples>.

6 *Giorgio Patrini*, *Francesco Cavalli* and *Henry Adjer*, 'The State of Deepfakes: Reality under Attack' (Deeptrace 2018) Annual Report v2.3 2.

7 *Sensity* was formerly known as *Deeptrace*, in their report they mentioned when the term was first used: "The term deepfake was first coined by the Reddit user u/deepfakes, who created a Reddit forum of the same name on November 2nd 2017. This forum was dedicated to the creation and use of deep learning software for synthetically faceswapping female celebrities into pornographic videos. Since Reddit's removal of /r/Deepfakes on February 7th 2018, deepfakes have become increasingly commodified as new deepfake forums, tools, and services have emerged", see 'The State of Deepfakes 2019. Landscape, Threats, and Impact' (2019) <https://sensity.ai/reports/>. See also for a comprehensive technical background on the state of deep fakes *European Parliamentary Research Service Scientific Foresight Unit* (STOA), 'Tackling Deepfakes in European Policy' (EU Publications Office 2021) PE 690.039.

The company also expects these numbers to double every six months starting in 2020, which is a significant trend that confirms the relevance of addressing deep fake phenomena.

The increased commodification of deep fakes through apps, tools, services, and forums is becoming an issue for regulators, because it has enabled accessibility of these tools and programs to a non-technical audience. For example, an investigation into the messaging platform *Telegram* revealed a new ecosystem of deep fakes relying on bots that provide free and simple user interface through smartphones or computers.[8] The user only has to upload an image to the bot and receive the processed image, or deep fake, after a short while. This increased accessibility has resulted in a mainstream effect of deep fake technology which is becoming popular for online users, such as for example the use of face swap apps on social media. In particular, regulators are increasingly concerned about the development and application of this technology in two areas: image-based sexual abuse and disinformation.

First, it is essential to highlight that deep fakes disproportionately affect women and queer communities.[9] For instance, 96% of deep fakes are pornographic content where women constitute 90 % of the target leading to reputational harms, invasion of privacy, and harassment, among the most common harms. Therefore, there is a gendered dimension to deep fake phenomena that must not be forgotten when considering regulating this technology. As *Citron* points out when analyzing sexual privacy issues, deep fake technology is often weaponized against women to terrify and silence them. As a consequence, deep fakes take women's agency over their bodies and make it difficult to stay online, get or keep a job, and feel safe.[10]

Second, even though only around 4% of deep fakes are non-pornographic,[11] they can easily be weaponized to become a systemic risk if, for example, they compromise electoral integrity or undermine democratic processes. For example, in 2019 *Future Advocacy* and UK artist *Bill Poster* released a deep fake where *Boris Johnson* and his rival *Jeremy Corbyn* endorsed each other for prime minister to illustrate the potential of deep fakes to undermine democracy just a month away from the UK General Election.[12] Most regulatory actions have been catalyzed by the looming threat of these types of deep fakes, even if they represent a fraction of the total amount of documented cases.

Though legal evidence tampering and fraud schemes are also increasingly becoming areas where deep fakes are used with criminal or national security implications, they are not yet quite as popular as the use of deep fakes for producing fake non-consensual pornogra-

8 *Giorgio Patrini*, 'Automating Image Abuse: Deepfake Bots on Telegram' (20 October 2020) <https://giorgiop.github.io/posts/2020/10/20/automating-image-abuse/>.

9 *Mary Anne Franks* and *Ari Waldman*, 'Sex, Lies, and Videotape: Deep Fakes and Free Speech Delusions' (2019) 78 Md. L. Rev. 892.

10 See *Danielle Citron*, 'Sexual Privacy' (University of Maryland Faculty Scholarship 2018) <https://digitalcommons.law.umaryland.edu/fac_pubs/1600>.

11 'The State of Deepfakes 2019. Landscape, Threats, and Impact' (n 7).

12 *Areeq Chowdhury*, 'Deepfakes – Press Release' (Future Advocacy, 12 November 2019) <https://futureadvocacy.com/deepfakes-press-release/>.

phy or disinformation. This regulatory concern has drastically increased as experts worry about the stunning evolution of the quality of the deep fakes produced, making it more difficult to detect for the average user whether there has been audiovisual manipulation or not.

Considering the rapid development of deep fake technology and the proliferation of harms, this article aims to analyze the EU level approach to deep fakes in relation to AI regulation. This choice is motivated by the inclusion of deep fakes in the European Commission Proposal for a Regulation laying down harmonised rules on Artificial Intelligence (hereinafter: AI Act),[13] and the nature of the provisions that are applicable to deep fake technology within the Act. The first section will provide a brief overview of deep fake phenomena, the current regulatory context and how it came it intersects with the AI Act. The second section will analyze the issues and challenges of adopting a legal definition for deep fakes to highlight consensus and differences among scholars, institutions, and industry players. Getting the scope of the definition right is essential to address appropriately the different harms stemming from deep fake technology, specifically in regards to disinformation and image-based sexual abuse. The third section focuses on the transparency requirements for deep fakes under the EU Artificial Intelligence Act proposal. This obligation will be examined in light of disclosure and labelling obligations already tested in disinformation strategies, in particular in the implementation of the EU Code of Practice on Disinformation. Finally, this contribution will provide on its fourth section a summary comparative view on other jurisdiction responses to deep fakes, mainly in the United States (US) and the United Kingdom (UK) in contrast to the European approach.

2. Overview of current regulatory context

Already in 2018, the European Commission and the European External Action Service were concerned by audiovisual (AV) manipulation similar to deep fakes without explicitly using the word "deep fakes" in their strategy to counter disinformation in view of the EU Parliamentary elections of 2019.[14] More recently, the European Parliament raised strong concerns about the proliferation of deep fakes, and decided to commission a study on deep fakes. Other organizations such as Europol and United Nations have also studied the phenomena. In their research all of these institutions consider deep fakes as an emerging threat: "Coupled with the reach and speed of the Internet, social media and messaging applications, deepfakes can quickly reach millions of people in an extremely short period of time. Because of this, deepfakes have been identified as a powerful weapon in today's disinfor-

13 Proposal for a Regulation of the European Parliament and of the Council Laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. COM(2021) 206 final.

14 *European Commission*, 'Tackling Online Disinformation: A European Approach' COM (2018) 236 Final.

mation wars, whereby people can no longer rely on what they see or hear".[15] Furthermore, the confirmation of foreign interference in the 2016 US presidential election prompted regulatory initiatives under the cybersecurity aegis. It is no secret that AI has amplified the phenomenon of disinformation, and deep fakes are becoming a powerful tool of that strategy.

Deep fake technology lies at the intersection of different possible regulatory frameworks providing an interesting case to explore the regulatory challenges of AI in the context of the European Union (EU). On the one hand, it is possible to consider that malicious uses of deep fakes are covered under some criminal statute when it is used for fraud schemes, as in the *Euler Hermes* example mentioned before, or when it comes to legal evidence tampering. However, in the insurance fraud example, allegedly because they could not identify the culprit there was no criminal case. Online anonymity, the high burden of proof and the broader accessibility of the technology are some of the reasons why deep fakes are challenging cases to bring to court. On the other hand, these examples are not the main malicious uses of deep fakes, and therefore, for the majority of cases that involve the use of deep fake technology for fake non-consensual pornography or disinformation lawyers often struggle to make a case. A patchwork of provisions from data protection and privacy frameworks, intellectual property laws, defamation laws and even, platform liability regimes need to be interpreted to formulate the best defense, with the caveat that success rate of this type of cases is low, for the already mentioned reasons.

Carrie Goldberg is a US lawyer with one of the first practices specialized in online stalking, revenge porn and gendered violence cases. When asked about her firm tackling the growing threat of deep fakes, she admits that there is no right law to prosecute whoever is responsible for the deep fake creation, although one could image defamation laws to be useful in these cases since something that is false is presented as true. However, because victims see their online sexual privacy violated, lawyer's strategies focus on content take downs rather than prosecution in court. More importantly, she questions the role of online platforms in disseminating deep fakes: "Why does Google even have deepfakes coming up within their algorithms? The weighting of that is a core issue".[16] Her answer points to an important regulatory issue which is currently under development in the EU regarding platform liability regimes.[17] Since one of the major issues, as it will be explained later on in

15 *United Nations Interregional Crime and Justice Research Institute* (UNICRI), 'Algorithms and Terrorism: The Malicious Use of Artificial Intelligence for Terrorist Purposes' (United Nations Office of Counter-Terrorism (UNOCT) 2021) <http://unicri.it/News/Algorithms-Terrorism-Malicious-Use-Artificial-Intelligence-Terrorist-Purposes>.

16 *Huck Magazine*, 'How Carrie Goldberg Turned Litigation into an Act of Protest' (20 August 2019) <https://www.huckmag.com/art-and-culture/books-art-and-culture/how-carrie-goldberg-turned-litigation-into-activism/>.

17 See *Carsten Ullrich*, 'Unlawful Content Online: Towards A New Regulatory Framework For Online Platforms' (Nomos 2021). And see *Mark Cole*, *Christina Etteldorf* and *Carsten Ullrich*, 'Cross-Border Dissemination of Online Content – Current and Possible Future Regulation of the Online Environment with a Focus on the EU E-Commerce Directive (Open Access)' (Nomos 2020).

this article, is the lack of applicable laws that can effectively offer a remedy to victims of deep fakes. Therefore, it could be imagined that deep fakes could come into play when considering updating rules that concern the main distribution channels for deep fakes. More precisely, it could be expected that in the process of amending the e-Commerce Directive online abuses related to deep fakes will be discussed. However, as pointed out by Member of the European Parliament *Karen Melchior*[18] in an event on tackling gender based online violence, deep fakes do not necessarily fit with the core reasons and the horizontal approach of the proposed Digital Services Act (hereinafter: DSA).[19] Deep fakes are above all the result of an AI system; thus, a more sectorial approach was deemed necessary when considering regulating them. Therefore, deep fakes were included in the AI Act raising normative questions on the European approach to this particular phenomenon.

In this regulatory context characterized by a plurality of possibly applicable provisions from laws and legislative proposals, the EU approach to deep fakes must be analyzed against the background of its strategy against disinformation and its AI regulation strategy.

3. Summary of the proposed AI Act

The European Commission published the proposed Artificial Intelligence Act in April 2021. This proposal follows the "White Paper on AI – A European approach to excellence and trust" [20] published in February 2020, in which the European Commission already discussed policy options on addressing risks while promoting benefits of AI technology. However, regulating AI systems has been on the workings since 2018 with the High-Level Expert Group on Artificial Intelligence, which started a discussion on the shape of a possible AI regulatory intervention in the EU.[21] The AI Act codifies principles for creating a trustworthy AI model respectful of democratic values, human rights and the rule of law. The proposed Act is currently under review by the European Parliament and the Council of the European Union. It is likely that at least three to four years will pass before an agreed version of the proposed regulation will enter into force.

The proposed horizontal legislation endorses a risk-based approach to AI, classifying systems based on their level of risk with differentiated obligations for each level. It propos-

18 In this event *BBC* journalist *Marianna Spring* interviewed MEP *Karen Melchior*, and advocacy organisations representatives. The event was organized in behalf of the MEP by *AWO Agency*. 'Tackling Gender Based Online Violence in the Digital Services Act' (Online Event, 30 September 2021).

19 Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC, COM (2020) 825 final.

20 *European Commission*, 'White Paper "On Artificial Intelligence – A European Approach to Excellence and Trust"' COM(2020) 65 final.

21 See *European Commission*, 'Commission Appoints Expert Group on AI and Launches the European AI Alliance' (Shaping Europe's digital future, 14 June 2018) <https://digital-strategy.ec.europa.eu/en/news/commission-appoints-expert-group-ai-and-launches-european-ai-alliance>.

es classifying AI systems in four categories: unacceptable risks (Art. 5), high-risk (Art. 6), AI with specific transparency obligations or limited risk (Art. 52) and minimal or no risk AI (Art. 69). For the higher risk category, regulation is stricter and includes audits and regular compliance and monitoring requirements. Most of the Act is devoted to these high-risk profile AI systems and their obligations. Annex III of the proposed AI Act lists the systems considered as high-risk. This, however, does not mean that AI systems that are not explicitly discussed are not under the scope of the regulation. This categorization can be reviewed by the European Commission (Art. 67) and thus, systems can be subject to a re-assessment of their risk procedure.

In the lower category risk the Act proposes to manage risk through non-binding self-regulatory instruments such as codes of conduct. Deep fakes were introduced in the AI Act as AI systems with specific transparency obligations or limited risk. This risk classification is considered subjective and could be challenged when it comes to deep fake technology which, as already mentioned, has a dual use nature. How explicitly these systems were mentioned and fall under the scope of the AI Act will be discussed in detail on section III. of this article.

The Act will be applicable to providers, importers and distributors as well as users of AI systems inside of the EU. However, as it aims to be one of the first-ever comprehensive frameworks regulating AI it is anticipated to set a global standard.

Among the proposal's key elements that have stirred debate is the material scope of the proposed regulation and the enforcement mechanisms. First, on the material scope, *Hildebrandt,* among other scholars, considers that the European Commission proposed a broad definition of AI systems.[22] This definition was inspired by the *Organisation for Economic Co-operation and Development* (OECD) definition, that includes machine learning, expert systems, and statistical models. The inclusion or not of different AI techniques under the definition for AI systems is a major point of concern for scholars and commentators.[23] Second, the architecture of enforcement of the proposed AI Act has been the object of criticism. In particular, the creation of a new European Artificial Intelligence Board as an oversight mechanism which has been compared to the General Data Protection Regulation (GDPR)[24] type of enforcement which lately has been under scrutiny for its results. Both of these issues will be addressed in relation to deep fakes later on in this article.

22 *Mireille Hildebrandt*, 'A Commentary on the Proposal for an EU AI Act of 21 April 2021' (Vrije Universiteit Brussel, 19 July 2021) 1.

23 *Virginia Dignum* and *Catelijne Muller*, 'Artificial Intelligence Act: Analysis and Recommendations' (ALLAI 2021) 9.

24 Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation) [2016] OJ L 119/1.

II. Defining deep fakes: issues and challenges

1. Building consensus for a legal definition of deep fakes

Deep fakes are described in Art. 52 (3) of the proposed AI Act as an "AI system that generates or manipulates image, audio or video content that appreciably resembles existing persons, objects, places or other entities or events and would falsely appear to a person to be authentic or truthful ('deep fake')". Although the Commission did not explicitly offer an official definition, under Art. 3 of the AI Act proposal, they refer to deep fakes in line with the broad consensual definition of what constitutes a deep fake. This section will discuss the different perspectives on what constitutes a deep fake and how they fit into the description in the AI Act proposal.

a) Scholarly perspectives

The term deep fakes was first identified in 2018 on a *Reddit* community post and, since then, has gained mainstream visibility. The term deep fake,[25] which evolves from the words 'deep learning' and 'fake', does not have a rigorous definition in the same way there is not one for 'fake news'.[26] This vagueness in the definition makes it harder for policymakers and users to identify and agree on what is meant by using the term deep fakes compared to other types of audiovisual manipulation that do not use artificial intelligence (i.e., *cheap fakes, shallowfakes*). It also highlights the difficulties of delimiting the subject matter of a possible regulatory approach.

From legal scholarship, *Citron* and *Chesney* were the first to survey harms and potential responses to deep fakes comprehensively. They defined them broadly as "the full range of hyper-realistic digital falsification of images, video, and audio".[27] Social scientists, such as *Paris* and *Donovan*, also adhere to a broad definition, where the term deep fakes refers to the "use some form of 'deep' or machine learning to hybridize or generate human bodies and faces".[28] Deep fakes are on the sophisticated end of the audiovisual media manipula-

25 "The term originally came from a Reddit user called 'deepfakes', who, in December 2017, used off-the-shelf AI tools to paste celebrities' faces onto pornographic video clips. The username was simply a portmanteau of 'deep learning' (the particular flavor of AI used for the task) and 'fakes'. Although the term was originally only applied to pornographic fakes, it was quickly adopted as shorthand for a broad range of video and imagery edited using machine learning". See *Patrini*, *Cavalli* and *Adjer* (n 6) 3.

26 For more on the typology and different concepts related to 'fake news' see *Claire Wardle* and *Hossein Derakhshan*, 'Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making' (Council of Europe 2017) DGI(2017)09.

27 *Robert Chesney* and *Danielle Keats Citron*, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security' (2019) 107 California Law Review 1753.

28 *Britt Paris* and *Joan Donovan*, 'Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence' (Data & Society, 2019) 5.

tion spectrum. To distinguish deep fakes from other AV manipulation techniques, they coined the term *cheap fakes*, which will be discussed later in this section.

From a cybersecurity and forensic experts' perspective, deep fakes are considered exclusively through the lense of deep learning. As *Hany Farid*, one of the world's foremost experts in deep fake detection explains, "at its core a deep fake is a fake image, video or audio that has been synthesized by a computer with very little human intervention in it".[29] From this perspective, the complete automation of creating a fake image, video, or audio makes it different from other traditional forms of faking content. It also democratizes its creation process since almost no skill is required. Moreover, the average internet user has access to software, websites, and apps, allowing them to create good quality deep fakes, moving away from the technical expertise once exclusive to movie studios or computer science experts.[30]

Furthermore, defining deep fakes has a further layer of complexity, which is that under the umbrella of deep fakes several methods can be included. For example, "face swaps, audio deepfakes (copying someone's voice), deepfake puppetry or facial reenactment (mapping a targets face to an actor's and manipulating it like that), and deepfake lip-synching (created video of someone speaking from audio and footage of their face)".[31] These are some of the popular variations of deep fakes. Currently, the term is also used to describe synthetic media applications and new creations like *StyleGAN*, which creates realistic images of people that do not exist.[32] The malicious use of this technique was showcased in the "Katie Jones" case, whose profile on *LinkedIn* claimed to work at the *Center for Strategic and International Studies*, but is thought to be a deep fake created for a foreign spying operation.[33]

29 *The Takeaway*, 'Are Deepfakes the Next Fake News?' <https://www.wnycstudios.org/podcasts/takeaway/segments/deepfakes-fake-news-artificial-intelligence>.

30 There is broad range of accessible off-the-shelf software to create deep fakes such *DeepFaceLab3*, *FaceApp* and other public repositories. For a survey on available repositories see: *Thanh Thi Nguyen* and others, 'Deep Learning for Deepfakes Creation and Detection: A Survey' (2021) arXiv:1909.11573 [cs, eess] <http://arxiv.org/abs/1909.11573>.

31 *James Vincent*, 'Why We Need a Better Definition of "Deepfake"' (The Verge, 22 May 2018) <https://www.theverge.com/2018/5/22/17380306/deepfake-definition-ai-manipulation-fake-news>.

32 *StyleGAN* was introduced by *Nvidia* researchers in 2018 as a technique to create with AI fake human faces. There is some debate whether *StyleGAN* should be considered under the term deep fake since it creates still images. However, since it is open source, the results are realistic enough to deceive, and the techniques can be applied to video it is assimilated to the deep fake term. See *Synced*, 'NVIDIA Open-Sources Hyper-Realistic Face Generator StyleGAN' (SyncedReview, 9 February 2019) <https://medium.com/syncedreview/nvidia-open-sources-hyper-realistic-face-generator-stylegan-f346e1a73826>.

33 *James Vincent*, 'A Spy Reportedly Used an AI-Generated Profile Picture to Connect with Sources on LinkedIn' (The Verge, 13 June 2019) <https://www.theverge.com/2019/6/13/18677341/ai-generated-fake-faces-spy-linked-in-contacts-associated-press>.

b) Industry stakeholders' perspectives

For tech companies working on deep fake creation or detection the definition of deep fakes is varied and mostly narrower in scope than the one provided by scholars. In 2018 *Deeptrace* provided one of the first compact definitions in the market: "deepfakes refer to any photo-realistic audiovisual content produced with the aid of deep learning. It also refers to the technology creating it. The term implies its misuse for illicit or unethical purposes."[34] In the same line of thought, several major platforms such as *Twitter*, *Facebook*, *Pornhub*, *TikTok*, and *YouTube* have progressively, and somewhat reluctantly, implemented policies between 2018 and 2020 to ban *manipulated, synthetic, or altered* media on their sites.[35] To delimit the scope of their policies, they each crafted their own definition of deep fakes.

One extensively discussed example in media outlets was *Facebook*'s ban of deep fakes in January 2020. Their policy was a response to the pressures to counter disinformation threats given the 2020 US presidential election, which brought to the forefront the debates over deep fakes in the context of disinformation and the state-of-the-art technology needed to counter the threat effectively.[36] The policy provided two criteria to determine whether a video could be labelled as a deep fake and thus, subject to the ban on their site. According to *Facebook*, deep fakes are: (i) the product of AI or machine learning, and (ii) they intend to "mislead someone into thinking that a subject of the video said words that they did not actually say".[37] This definition seems in line with the mainstream criteria agreed by scholars and practitioners until then.

However, the reactions to this policy definition were mixed. Some of the main critics were concerned with the wording in the criteria, meaning whether the definition of deep fakes only included speech or if a video of someone doing something they did not do will

34 *Patrini*, *Cavalli* and *Adjer* (n 6).

35 The report by *Democracy Reporting International* compiles most of the main social media platforms policies towards deep fakes. Moreover, their research shows that private messaging apps (i.e. *WhatsApp*, *Telegram*, *Facebook Messenger*) do not have any relevant detection or enforcement policy concerning deep fakes. See *Madeline Brady*, 'Deepfakes: A New Disinformation Threat' (Democracy Reporting International, 2020) 23 <https://democracy-reporting.org/wp-content/uploads/2020/08/2020-09-01-DRI-deepfake-publication-no-1.pdf>.

36 *The Guardian,* 'Facebook Bans "deepfake" Videos in Run-up to US Election' (7 January 2020) <http://www.theguardian.com/technology/2020/jan/07/facebook-bans-deepfake-videos-in-run-up-to-us-election>.

37 The criteria for a video post to be considered a deep fake under *Facebook*'s Media Manipulation Policy is as follows: "It has been edited or synthesized – beyond adjustments for clarity or quality – in ways that aren't apparent to an average person and would likely mislead someone into thinking that a subject of the video said words that they did not actually say. And: It is the product of artificial intelligence or machine learning that merges, replaces or superimposes content onto a video, making it appear to be authentic." See *Facebook*, 'Enforcing Against Manipulated Media' (About Facebook, 7 January 2020) <https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/>.

also classify as a deep fake.[38] *Citron*, who advised *Facebook* on this policy, considered it a positive step, though, as she pointed out: "Some deep fakes don't involve words, just actions like deep fake sex videos. They invade sexual privacy and cause profound harm".[39] Therefore, videos showing people doing things they never did should also be included. Another critical point of criticism was the fact that this policy excluded *cheap fakes*, which account for most of the visual disinformation on the platform.[40] For example, disingenuous doctoring of videos during election period "will go entirely untouched by the new 'tougher' policy."[41] For most commentators and scholars, it seems as *Facebook* carved a very narrow definition of deep fakes in contrast to the consensus.

A different example is the case of *Twitter*. The platform carefully included deep fakes and synthetic media in its policy. The criterion for labelling content do not focus on whether AI has been used, but "whether media have been significantly and deceptively altered or fabricated".[42] Therefore, the broad scope of their policy includes *cheap fakes* that can be dangerous in the context of electoral disinformation and allows for an expeditious action according to their guidelines. Additionally, their policy details the different actions to be taken according to a set criterion, which addresses trust and transparency concerns.

c) Elements of consensus

Generally, academia and practitioners have a broad consensus that a deep fake is commonly understood as having two differentiating elements. First, it is based on AI techniques, more precisely on deep learning to automate some parts of the audiovisual editing process. Second, deep fakes have the potential to deceive, and perhaps meaningfully affect lives.[43]

Scholars and practitioners alike agree that there is an intent requirement in creating a deep fake that needs to be considered when determining regulatory choices. In particular, the creation process of convincingly realistic-looking deep fakes, despite the *hype* and the increasing commodification of the technical solutions, requires a considerable amount of "intentional" work as explained by *Burkell* and *Gosse*. They stress the idea that the algorithm underlying the creation of a deep fake is not going to create a non-consensual fake

38 *Hadas Gold*, 'Facebook Tries to Curb Deepfake Videos as 2020 Election Heats Up' (CNN, 7 January 2020) <https://edition.cnn.com/2020/01/07/tech/facebook-deepfake-video-policy/index.html>.

39 *Robert Chesney*, 'Facebook Takes a Step Forward on Deepfakes – And Stumbles' (Lawfare, 8 January 2020) <https://www.lawfareblog.com/facebook-takes-step-forward-deepfakes-and-stumbles>; 'Facebook Tries to Curb Deepfake Videos as 2020 Election Heats Up' (n 38).

40 *Gilad Edelman*, 'Facebook's Deepfake Ban Is a Solution to a Distant Problem' (Wired, 7 January 2020) <https://www.wired.com/story/facebook-deepfake-ban-disinformation/>.

41 *Natasha Lomas*, 'Facebook Bans Deceptive Deepfakes and Some Misleadingly Modified Media' (TechCrunch, 7 January 2020) <https://social.techcrunch.com/2020/01/07/facebook-bans-deceptive-deepfakes-and-some-misleadingly-modified-media/>.

42 *Twitter*, 'Our Synthetic and Manipulated Media Policy' (Twitter Help) <https://help.twitter.com/en/rules-and-policies/manipulated-media>.

43 *Vincent* (n 31).

pornography content unless it is programmed to do so.[44] This intention requirement is also applicable to the creation of disinformation deep fakes. As argued by technology philosophers, deep fakes are not inherently deceptive.[45] The deceptive effect of deep fakes is the by-product of the creator's intent and the conditions in which it is released.

2. Challenges to a practical definition

Although, these two defining elements seem to appear simple, building a concise legal definition for a regulation is not straightforward and will present practical challenges in its implementation. In particular, there are two difficulties to overcome to get a clear picture of why defining deep fakes has become an issue.

a) Setting the boundary between deep fakes and other forms of audiovisual manipulation

First, a crucial difficulty in understanding what deep fakes are comes from the misuse of the term. The term deep fake has been increasingly misused to refer to audiovisual content manipulation, which does not rely on AI and is sometimes referred to as *cheap fakes.* Drawing the line between cheap fakes and deep fakes is tricky because of the technical difficulties of detecting if AI has been used to modify a video. As discussed by *Britt Paris* and *Joan Donovan*, there is a spectrum between cheap fakes and deep fakes.[46] The ability to manipulate audiovisual content is not new. However, the speed at which audiovisual content can be manipulated, the unprecedented quality of the result, and the availability to be used off-the-shelf by unskilled users are the new variables. Knowing where to draw the line in this spectrum between cheap fakes and deep fakes is likely to be the critical factor in regulating AV manipulation because of the lack of consensus on the practical implementation of such policies.

This difficult call has important implications for policymakers and regulators, as it showed the high-profile case of a video of US House of Representatives Speaker *Nancy Pelosi* on *Facebook*. In that case, the video was digitally slowed down to make it appear as if she was drunk and slurring her words. Technically, it was a classic AV manipulation case that did not fit the definition of deep fakes since no AI was involved. For this reason, *Facebook* policy against manipulated media could not be enforced. The unwillingness of *Facebook* to take down a video that harmed a politician's reputation and had a pervasive disin-

44 *Jacquelyn Burkell* and *Chandell Gosse*, 'Nothing New Here: Emphasizing the Social and Cultural Context of Deepfakes' (2019) First Monday 24(12) <https://firstmonday.org/ojs/index.php/fm/article/view/10287>.

45 *Adrienne Ruiter*, 'The Distinct Wrong of Deepfakes' (2021) Philosophy & Technology 9. See more generally on the nature of AI deception *Luciano Floridi*, 'Artificial Intelligence, Deepfakes and a Future of Ectypes' (2018) 31 Philosophy & Technology 317.

46 *Paris* and *Donovan* (n 28).

formation effect, in an already tense political context in the US,[47] illustrates how important it is to consider the harms caused by lower-tech media manipulation in the discussion on regulating deep fakes. In this case, *Facebook*'s response was to label the video, which is the preferred regulatory option being discussed in the context of the EU AI Act proposal. However, in that case, labelling was considered an insufficient and ineffective response by media commentators, fact-checking organizations and advocacy groups. In particular, because labelling relies on fact-checking, and this type of measure often overlooks the problem of content going viral before it can be debunked.[48]

b) The "Liar's dividends" effects of deep fakes

Second, and more challenging, the term deep fake is also being co-opted to discredit audiovisual content and cast doubt on the veracity of video content presented as evidence. This has been particularly problematic for human rights activists that rely on video footage to document abuses or illegal activities.[49] For example, there is a real danger that authoritarian regimes use deep fakes to cast doubt on the veracity of human right activist accusations as it was the case this year in Myanmar. A video that was broadcasted by *Myawaddy TV*, which is a television network owned by Myanmar's military, showed the former chief minister confessing to bribing ousted leader *Aung San Suu Kyi*.[50] Although there was no confirmation as to whether the video was a deep fake, the broadcasting was followed by riots. As *Gregory* points out, "one big harm in contexts like Myanmar is the ability to undermine a true video by claiming it's false, and the way manipulated media can be used to inflame religious and ethnic conflicts."[51] Deep fakes are also problematic in highly polarized political contexts. For example, following the US riots in Capitol Hill, a video circulated where ex-president *Trump* conceded victory to president *Joe Biden*. The video was considered a deep fake vastly on social media networks even though no credible source could detect audiovisual manipulation.[52] This episode created confusion, in a critical post-election period, among all types of media outlets. *Chesney* and *Citron* have warned against the "liar's dividend" effects of deep fakes. Deep fakes make it "easier for liars to avoid accountability

47 *The Guardian*, 'Facebook Refuses to Remove Doctored Nancy Pelosi Video' (3 August 2020) <http://www.theguardian.com/us-news/2020/aug/03/facebook-fake-nancy-pelosi-video-false-label>.

48 *Edelman* (n 40).

49 See *WITNESS*, 'Prepare, Don't Panic: Synthetic Media and Deepfakes' (WITNESS Media Lab) <https://lab.witness.org/projects/synthetic-media-and-deep-fakes/>.

50 *KrASIA*, 'Did Myanmar's Military Deepfake a Minister's Corruption Confession?' (24 March 2021) <https://kr-asia.com/did-myanmars-military-deepfake-a-ministers-corruption-confession>.

51 ibid.

52 *Reuters*, 'Fact Check: Donald Trump Concession Video Not a "Confirmed Deepfake"' (11 January 2021) <https://www.reuters.com/article/uk-factcheck-trump-consession-video-deep-idUSKBN29G2NL>.

for things that are in fact true."[53] Consequently, they erode trust in society. In the long term, the recurrent misuse of the term creates an information ecosystem of mistrust.

Because of the practical challenges of determining when audiovisual media has been manipulated using AI and the societal hurdles of understanding what constitutes a deep fake, the AI Act proposal should consider including an explicit definition of what is regulated as a deep fake. Thereby, cementing consensus among the community and clarifying the subject matter of the regulation to ensure its enforceability. Moreover, if the objective is to curtail digital disinformation significantly, the legislator needs to consider measures to fill the gap in a continuum between both ends of lower audiovisual manipulation and sophisticated AI-based manipulation.

3. Further elements of analysis

a) A technology driven approach to deep fake regulation

As already mentioned, some of the narrower definitions of deep fakes focus on the use of AI or machine learning to produce deep fakes as one of their main criteria. This narrow interpretation translates into a high threshold requirement of proof. Detecting whether AI has been used to create the content in question requires state-of-the-art software and is a resource-intensive task that is not yet available to everyone. Even fact-checking deep fakes is a demanding task that requires specialized software and knowledge in forensics and, thus, is not so easily viable.[54]

Scholars such a *Farid* warn that forensic technology capabilities detecting a real from a fake are decades away of being conclusive, in particular due to the adversarial nature by which deep fakes are created.[55] Although governments are presumed to be developing technologies that can detect doctored images and audio to improve media platform forensics, it is unknown to what extent they have achieved success. In the EU, working closely with platforms has been the preferred option for enhancing the detection of different types of illegal content in the online ecosystem, as demonstrated by the different self-regulatory schemes (i.e., codes of conduct) implemented or encouraged by the European Commission in recent years. However, for example, *Facebook* has already admitted that they do not currently have the technology to stay ahead of adversarial media manipulation and that more

53 *Chesney* and *Citron* (n 27) 1758.

54 Moreover, scaling the methods to detect deep fakes is often cost prohibitive since it implies producing hundreds of thousands of deep fakes videos to train de computer vision or multimodal models. See *Brian Dolhansky* and others, 'The DeepFake Detection Challenge (DFDC) Dataset' (2020) arXiv:2006.07397 [cs] <http://arxiv.org/abs/2006.07397>.

55 See *Chesney and Citron (n 27) 1788*, quoting *Prof. Hany Farid* on the state of deep fake forensic technology.

innovation in this area is needed.[56] To counter these threats, they have partnered with leading universities in the US to help them improve their image and video technology. These official statements by the company read along with their proposed research agenda implies that currently, automated detection of deep fakes on the platform is at its early stages.[57]

To improve deep fake detection techniques, *Facebook* and *Google* launched in 2019 the Deepfake Detection Challenge (DFDC), which made public to researchers an open-source database of deep fakes to help train and test automated detection tools.[58] Even though on a technical level, there is no guarantee that even by allowing access to data, the algorithms will be effective since data can be contaminated by a technique known as adversarial machine learning, further skewing, for example, biases in the training data. The results of the project confirmed the need for more research on deep fake detection and media forensics since in real-life conditions, "the top model achieved an accuracy of 65.18 percent."[59] This result became the new shared baseline for the AI community working on this issue. However, for the purpose of protecting online users from deep fakes threats, relying solely on detection at a 65% accuracy rate does not seem a high enough level yet. As pointed out by researchers, "the DFDC results also show that this is still very much an unsolved problem. None of the 2,114 participants, which included leading experts from around the globe, achieved 70 percent accuracy on unseen deepfakes in the black box dataset."[60] These results also confirm that measures beyond technical self-regulatory initiatives are needed.

An overly emphasis on detection as a mean to counter deep fakes biases self-regulation efforts by reducing the range of actions or "remedies" available to private actors to prevent the proliferation of malicious deep fakes. In particular, in a context where experts warn on the fleeting nature of getting the upper hand in adversarial network development, self-regulation efforts should not be reduced to a technology-driven solution. As argued by *Goldman*, a normative framework should move past the binary idea of remove-or-not remedy framework when there is enough evidence from platform governance studies that private actors implement around a dozen of other options when they are constraint to deal with un-

56 *Facebook*, 'Reports on Implementation of the Code of Practice on Disinformation' (2019) Section V.

57 *Will Knight*, 'Facebook Is Making Its Own AI Deepfakes to Head off a Disinformation Disaster' (2019) MIT Technology Review <https://www.technologyreview.com/2019/09/05/65353/facebook-is-making-ai-deepfakes-to-head-off-a-disinformation-disaster/>.

58 *MetaAI*, 'Deepfake Detection Challenge Results: An open initiative to advance AI' (12 June 2020) <https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/>.

59 The Deepfake Challenge showed that the highest performing models can achieve 82.56 percent accuracy if tested on the public dataset available to develop the models. However, if the same model was tested against a black box dataset, so against unforeseen examples, then the accuracy of the model decreases significantly to 65 percent. These results reinforce the importance of learning when developing detection methods so that they can be really useful in real life conditions. See more on Deepfake Challenge ibid.

60 ibid.

lawful or harmful content on their platforms.[61] Therefore, legislation on deep fakes should move past the idea that detection or not-detection are the only possible avenues for private actors to guarantee enforcement of applicable rules or prevent dissemination of deep fakes.

b) A distinct profile harm for deep fakes

Focusing on the technology used as the main criterion to determine if an audio or video is a deep fake can sidetrack legislators' and policymakers' attention on addressing the harms. In non-consensual deep fakes, as pointed out by *Citron* and *Franks*, the sharing of realistic-looking fakes can be as harmful as the sharing of authentic, intimate images.[62] For victims, the fact that it is a fake is one of the elements which makes deep fakes precisely so harmful. It robs them of their autonomy and agency and causes almost the same harm as if the actual intimate image was distributed.

A similar concern is discussed in the case of disinformation deep fakes. As *Adjer* points out, "Deepfake videos don't even have to be that good, as long as the person is recognizable and the graphics are good enough for a viewer to identify the person and see they're doing or saying something."[63] Being part of a negative video or audio (whether it is the name, voice, or image that is associated) leaves an imprint on the subject. In some cases, the risk of defamation is almost the same whether the content is fake or real.[64]

For this reason, *Diakopoulos* and *Johnson* introduced the notion of "persona plagiarism", which in their view sets deep fakes harms apart from misattribution, libel, or slander, in particular in the context of electoral disinformation.[65] *De Ruiters* expands this idea to explain that people's image and voice are markers of the self.[66] Hence, digital representations of our voice and face are connected to our identity and thus, there is a need to protect them. She argues that portraying someone in ways they would be unwilling to be portrayed is the distinctive aspect of deep fakes that renders them morally wrong. For this reason, a right to

61 *Eric Goldman*, 'Content Moderation Remedies' (2021) 27 Michigan Technology Law Review 1, Santa Clara Univ. Legal Studies Research Paper <https://papers.ssrn.com/abstract=3810580>.

62 *Danielle Keats Citron* and *Mary Anne Franks*, 'Criminalizing Revenge Porn' (2014) 49 Wake Forestlaw Review 48.

63 *Meredith Somers*, 'Deepfakes, Explained' (MIT Sloan, 21 July 2020) <https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained>.

64 *Perot* and *Mostert* also analyses if persona protection and the right of publicity can be used to redress deep fakes, and found that in UK and the US, deep fakes could be addressed by existing regulation. However, they conclude that these causes of action do not address adequately the potential of deep fakes, in particular when it comes to non-consensual deep fakes. See *Emma Perot* and *Frederick Mostert*, 'Fake It Till You Make It: An Examination of the US and English Approaches to Persona Protection As Applied to Deepfakes on Social Media' (SSRN, 2020) ID 3537052 <https://papers.ssrn.com/abstract=3537052>.

65 *Nicholas Diakopoulos* and *Deborah Johnson*, 'Anticipating and Addressing the Ethical Implications of Deepfakes in the Context of Elections' (2019) New Media & Society <https://papers.ssrn.com/abstract=3474183>.

66 *Ruiter* (n 45).

digital self-representation is needed to impede others from manipulating people's digital data in hyper-realistic footage without their consent. Therefore, deep fakes have a specific harm profile intrinsically linked to the ownership rights of one's image and data, since to generate deep fakes, the software must have access to data sets. In principle, the larger the data set, the better the quality of the deep fakes. However, and as it will be discussed later, data protection laws are not sufficient to deal with deep fake phenomena, though they can assuage victims' harms.

Against these considerations, the description of deep fakes provided in the proposed AI Act fares well. First, it does contain the two main elements of consensus on what constitutes a deep fake discussed in section II. 3. of this article. On the one hand it implies that the type of AI systems to which the act is applicable includes systems or techniques that manipulate images, audio, or video content. Therefore, the technical requirement criterion is included. As previously discussed, however, this does not mean that the definition would not have to struggle with the challenges of distinguishing lower AV manipulation from AI AV manipulation. On the other hand, it also mentions the intent that a deep fake must have and the potential it holds to deceive someone stating that if the content appears false or untruthful, it can be said to be a deep fake.

Second, an earlier version of the AI Act included rightly the idea that content that "has been altered to be sufficiently realistic representations of existing persons, objects, places or other entities or events"[67] is enough to be included in the deep fake category and subject to the transparency requirements under Art. 5 (3) of the AI Act proposal. As previously pointed out, content that looks alike is harmful enough without it necessarily being real. The official proposal changes the wording to "appreciably resembles existing persons, objects, places or other entities or events and would falsely appear to a person to be authentic or truthful ('deep fake')"[68] meaning that the content does not need to be realistic but to look-alike enough, which lessens the burden of proof for victims.

III. An approach to regulating deep fakes in the proposed AI Act

Deep fakes highlight the capabilities of AI to deceive people easily. Humans tend to trust what they see or hear, and deep fakes will exploit this weakness.[69] Moreover, because most

67 Leaked AI Act, first available in POLITICO Pro (link unavailable), and republished in *Natasha Lomas*, 'EU Plan for Risk-Based AI Rules to Set Fines as High as 4% of Global Turnover, per Leaked Draft' (TechCrunch, 14 April 2021) <https://techcrunch.com/2021/04/14/eu-plan-for-risk-based-ai-rules-to-set-fines-as-high-as-4-of-global-turnover-per-leaked-draft/>.

68 Artificial Intelligence Act (n 13) Art. 52 (3).

69 See *Foer* and *Resnick* for a reflection on the potential effects deep fakes can have on our collective memories and sense of reality. *Franklin Foer*, 'The Era of Fake Video Begins' (The Atlantic, 8 April 2018) <https://www.theatlantic.com/magazine/archive/2018/05/realitys-end/556877/>; *Brian Resnick*, 'We're Underestimating the Mind-Warping Potential of Fake Video' (Vox, 20 April 2018) <https://www.vox.com/science-and-health/2018/4/20/17109764/deepfake-ai-false-memory-psychology-mandela-effect>.

deep fakes are visual content, they have the potential to reach a larger audience in terms of media dissemination which makes them dangerous: "Taken together, common cognitive biases and social media capabilities are behind the viral spread of falsehoods and decay of truth. […] Information cascades, natural attraction to negative and novel information, and filter bubbles provide an all-too-welcoming environment as deep-fake capacities mature and proliferate."[70] In this context, the question then is how to mitigate the risk of the malicious use of deep fakes while acknowledging the legitimate uses for this technology, such as, for example, creating video characters for e-learning platforms. The inclusion of deep fakes under the proposed AI Act can be interpreted as a first attempt to regulate some aspects of this technology in the EU. Therefore, the first question is whether the applicable provisions in the proposal are sufficient to address the evidenced harms of deep fakes, and therefore fulfill the stated objectives of the regulation.

1. Transparency requirements for deep fakes

First of all, it must be highlighted that one of the objects pursued by the AI Act in Art. 1 is to lay down transparency rules specifically for deep fakes, which are included in the broader *AI system* category. In line with Recital 70,

> "users, who use an AI system to generate or manipulate image, audio or video content that appreciably resembles existing persons, places or events and would falsely appear to a person to be authentic, should disclose that the content has been artificially created or manipulated by labelling the artificial intelligence output accordingly and disclosing its artificial origin."[71]

The general objective is to protect natural persons of the risks of impersonation or deception irrespective of whether the AI system qualifies as high-risk or not.

Furthermore, Art. 52 (3) in Title IV of the AI Act lays down transparency obligations for specific AI systems:

> "Users of an AI system that generates or manipulates image, audio or video content that appreciably resembles existing persons, objects, places or other entities or events and would falsely appear to a person to be authentic or truthful ('deep fake'), shall disclose that the content has been artificially generated or manipulated."

This article imposes an obligation to disclose deep fakes, which is a minimal transparency requirement. As discussed in the Impact Assessment Report of the AI Act proposal,[72] a la-

70 *Chesney* and *Citron* (n 27).

71 Proposal for a Regulation of the European Parliament and of the Council Laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. COM(2021) 206 final (n 13).

72 *European Commission*, 'Impact Assessment Accompanying the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts' SWD(2021) 84 final.

belling obligation regarding deep fakes is considered a proportionate measure since deep fakes are not considered a high-risk technology. In this regard, the case of deep fakes shares some similarities to chatbots, which have been more studied in legal scholarship, and for which disclosure obligations have also come up as one of the preferred solutions.

The requirement of disclosure for deep fakes stems from the fact that it is manipulated audiovisual content. As discussed above in section I., it alludes to its deceptive potential, which is a consensual criterion to define deep fakes among scholars and industry players. As it stands, all deep fakes, including legitimate uses, are subject to disclosure. That seems unnecessary given the increasing proliferation of deep fake technology uses ranging from face-swapping apps for social media to skin-voices for optimal gaming immersion.

According to the European Commission Impact Assessment of the AI Act, several of the policy options evaluated indicated that labelling deep fakes would not be an applicable rule if the deep fake was used for legitimate purposes.[73] Labelling deep fakes is only meant to prevent the risk of manipulation. That is the intention of the provision according to Recital 70 of the AI Act proposal. However, the official wording in the AI Act provision does not distinguish between legitimate purposes and malicious uses. Therefore, to clarify the proposed regulation, the legislator could define legitimate purposes and malicious uses. Alternatively, it could propose to use the term synthetic media, understood as "any form of media generated by AI, including video, audio or images, and text",[74] and which *per se* does not have a malicious connotation as deep fakes. Using a different term may be helpful to distinguish malicious deep fakes from legitimate uses of deep fake technology and facilitate categorizing content in view of applying rules. In that way, only deceitful deep fakes would need to be labelled while legitimate synthetic media creation are not.

2. Labelling requirements for deep fakes

Since disclosure and labelling appear as the preferred regulatory options in the proposed AI Act for deep fakes, a second line of inquiry to pursue should be to assess whether this measure works and if it will be enough to counter the malicious use of deep fakes. In that regard, it appears relevant to examine the strategies already in place to counter disinformation in online platforms which have implemented warning and labels as their primary measures. The premise behind this measure is that if users are provided an informative notice or a warning on a piece of content, they will make an informed decision about what to believe in regards to that content.[75] However, media studies show that systematically this has not been a very effective measure in countering disinformation. *Kaiser* et al. have been studying labelling in disinformation context and their conclusion is that there is not enough evidence

73 ibid 45.

74 *Brady* (n 35) 2.

75 This premise also considers that all users are informed, which is not the case for deep fakes since there is not enough awareness on this phenomena yet. For this reason, media literacy is an important component that should follow regulation.

that these type of misinformation warning are effective: "Study after study has shown minimal effects for common warning designs. […] [W]e found that many study participants didn't even notice typical warnings – and when they did, they ignored the notices. Platforms sometimes claim the warnings work, but the drips of data they've released are unconvincing."[76] Moreover, platforms are not innovating or experimenting enough with new designs as argued by *Kaiser* et al. The status quo for countering online disinformation are still contextual warnings which are less effective than other designs such as interstitial warnings. Given the disputed rates of effectiveness on labelling content, there is no indication that applied to persuasive content such as deep fakes this measure will be any more effective. Therefore, it is not enough to require transparency when content has been manipulated, but there should be some measure to verify it produces effective outcomes. In particular, if new research on disinformation gives hope that there are innovative ways to achieve effectiveness that are just not being implemented. For these reasons, the proposed AI Act should be more precise on monitoring measures for minimal transparency requirements and provide guidelines on labels.

One further interpretation of this disclosure obligation under Art. 52 (3) would be to consider that the disclosure refers to the labelling of meta data, since in the context of AI labelling data is a common practice. However, this interpretation will encounter the difficulties and challenges related to deep fake detection and the state of automated filtering, which has been previously discussed, and thus, be undermined by the practical challenges on its implementation. Furthermore, this interpretation does not seem coherent with the overall intention of Art. 52 of the proposed AI Act which is to make users aware that they are interacting with an AI system, so they can "make informed choices or step back from a given situation" as stated in the Explanatory Memorandum that accompanies the proposed AI Act.[77]

3. Lessons from the Code of Practice on Disinformation

Moreover, in the context of disinformation, transparency measures that are interpreted and enforced by private companies have not worked very well in the past, such as in the EU Code of Practice on Disinformation (hereinafter: Code of Practice). In 2018, the European Commission decided to promote a Code of Practice, which became the first of its kind, to regulate the disinformation phenomena by mobilizing the private sector to tackle disinformation practices on online platforms, and more largely on promoting ethical behaviors within the industry. Signatories of the Code of Practice had reporting duties to the European

76 *Ben Kaiser*, *Jonathan Mayer* and *J. Nathan Matias*, 'Warnings That Work: Combating Misinformation Without Deplatforming' (Lawfare, 23 July 2021) <https://www.lawfareblog.com/warnings-work-combating-misinformation-without-deplatforming>.

77 Para. 5.2.4 of the Explanatory Memorandum in Proposal for a Regulation of the European Parliament and of the Council Laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts. COM(2021) 206 final (n 13).

Commission, and five intermediate reports assessing their implementation of the Code of Practice were submitted between January and May 2019.[78] These reports allowed for a – until then – rare and insightful first-hand perspective on how some online platforms enforced the Code of Practice. The *European Regulators Group for Audiovisual Media Services* (ERGA) assessed the enforcement challenges encountered in the implementation of the Code of Practice based on these intermediate reports.

Deterring AV manipulation or addressing deep fake phenomena were not key commitments of the Code of Practice, nor were they mentioned. However, signatories committed to publicly disclosing and labelling political advertising content and "issue-based advertising". In this way, the Code of Practice tested similar transparency requirements as those in the AI Act proposal. Additionally, considering the implications of deepfake technology on distorting democratic discourse and election integrity, the assessment of the Code of Practice becomes relevant to understand the limitations and challenges of this type of measure. Among the findings of *ERGA*'s assessment on the Code of Practice, they confirmed that labelling was one of the preferred self-regulatory measures by industry stakeholders. Thus, consumers were supported primarily through labelling and links to additional information. However, in practice, the implementation was challenging, and the data in the reports submitted is incomplete and does not allow for comparison among stakeholders.[79] They also found that some platforms did not report on the content they labelled as "false" once fact-checked. Moreover, in five EU Member States, platforms covered by this study had no designated certified third-party fact-checker. This resulted in the unavailability of some of the features that alert users on the trustworthiness of a post even though the content was flagged as fake by users, such as fact-checked labels. Furthermore, in *Facebook*'s case, the 'Context' button, which could have helped when labels were not available, was not active in some countries. These are some of the examples documented by *ERGA* that show enforcement issues with minimal disclosing requirements to enhance transparency.

A strengthened Code of Practice is being drafted, which will build up in terms of monitoring the recommendation made by *ERGA*[80] and other stakeholders from the 2018 Code of Practice assessments and enforcement challenges. The new draft was discussed at the end

78 The initial signatories of the Code of Practice included *Facebook*, *Google*, *Twitter* and *Mozilla*. The trade association representing online platforms (EDIMA) and trade associations representing the advertising industry and advertisers (the *European Association of Communications Agencies* (EACA), *IAB Europe*, the *World Federation of Advertisers* (WFA), the *WFA*'s Belgian national association, and the *Union of Belgian Advertisers*). Additional signatories have subscribed in 2020, these include *Microsoft*, *TikTok* and the French, Czech, Polish and Danish national associations affiliated with *EACA*. See more on the initial setup of the Code of Practice on Disinformation 2018 at https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation.

79 *ERGA*, 'ERGA Report on Disinformation: Assessment of the Implementation of the Code of Practice' (European Regulators Group for Audiovisual Media Services, 2020).

80 *ERGA*, 'ERGA Recommendations for the New Code of Practice on Disinformation' (European Regulators Group for Audiovisual Media Services, 2021).

of 2021, and the final version is expected in March 2022.[81] It is likely that this new version will include deep fakes as a threat to the online ecosystem. So far, the European Commission has stated in its guidance document that the strengthened Code of Practice will provide a comprehensive coverage of the current and emerging forms of manipulative behaviours, including deep fakes.[82] Furthermore, signatories will deploy and publish relevant policies and lay down baseline elements and objectives to counter deep fakes. Most importantly, "[t]he strengthened Code should take into consideration the transparency obligations for AI systems that generate or manipulate content and the list of manipulative practices prohibited under the proposal for Artificial Intelligence Act".[83]

Moreover, the new Code of Practice will be within a co-regulatory framework instead of the 2018 Code, considered a self-regulatory initiative. In that regard, *Marsden* et al. argue that a co-regulatory scheme is the most suitable option to regulate disinformation in the European context. Considering the positive obligations of States to intervene to protect rights, they submit that a scenario with no possibility of restrictions on free-speech is not possible nor appealing in the EU context.[84] The new Code of Practice is expected to become a code of conduct under the proposed DSA and be instrumental in monitoring the AI Act. Even though the 2018 Code of Practice raised concerns about safeguarding freedom of expression, this new framework promises to safeguard better users' rights, democracy, and European human rights standards.

Moreover, in 2020 and 2021, the stakeholders involved in the Code of Practice participated in a disinformation-monitoring program to regulate misleading information around coronavirus and vaccines.[85] Reducing the visibility of false rated content through "false" or "partly-false" labels, user notifications, interstitial warnings, and content demotion practices was an essential strategy for most stakeholders. The effectiveness of these measures, particularly labelling disinformation content, is not yet available as the first phase monitoring report does not provide a comprehensive overview of all the signatories nor can the individual reports provided be compared to a baseline.

81 *EU Newsroom*, '2021 Vademecum of the Assembly of the Signatories of the Code of Practice on Disinformation' <https://ec.europa.eu/newsroom/repository/document/2021-27/Vademe-cum_2021_Code_poQcw5VoJg362zUKq6VL774dsFs_78161.pdf>.

82 *European Commission*, 'Guidance to Strengthen Code of Practice on Disinformation' COM (2021) 262 final.

83 ibid 12.

84 *Chris Marsden*, *Trisha Meyer* and *Ian Brown*, 'Platform Values and Democratic Elections: How Can the Law Regulate Digital Disinformation?' (2020) 36 Computer Law & Security Review 105373.

85 *European Commission*, 'First Baseline Reports – Fighting COVID-19 Disinformation Monitoring Programme' (Shaping Europe's digital future) <https://digital-strategy.ec.europa.eu/en/library/first-baseline-reports-fighting-covid-19-disinformation-monitoring-programme>.

4. Other normative concerns regarding deep fakes

a) Normative concerns over enforcement mechanisms

As suggested by *Veale* and *Borgesius*, the inclusion of deep fakes in the AI Act raises more questions than it solves.[86] These scholars highlight the issue of practical enforcement of the disclosure provision. The disclosure obligation falls on the users and not the providers of the AI system. Therefore, users are expected to understand and determine the consequences of deep fakes they create on their own. This seems as a very challenging judgment call for users and an even harder decision to monitor since it is impossible for an enforcement body to investigate undisclosed deep fakes.[87] According to their analysis, the enforcement system under the AI Act relies heavily on market surveillance authorities because the AI Act focuses on product regulation and takes after the New Legislative Framework regime. However, deep fakes are a complex issue far from the regular activities carried out by these bodies, that at the very least, will need to invest and develop media forensic experience. Perhaps the European Artificial Intelligence Board, foreseen in the AI Act, will issue guidelines on deep fake technology and refine the disclosure requirements to enhance enforcement.

Another way to look at it, and as it has been suggested, is that enforcement concerning deep fakes will be discussed within the new Code of Practice on Disinformation. Either way, the uncertainty on the enforcement scheme of an already minimal requirement might lead to weak enforcement at its best.

This ambiguous enforcement setup is also important when it comes to exceptions. The disclosure requirement in Art. 52 (3)

"shall not apply where the use is authorised by law to detect, prevent, investigate and prosecute criminal offences or it is necessary for the exercise of the right to freedom of expression and the right to freedom of the arts and sciences guaranteed in the Charter of Fundamental Rights of the EU, and subject to appropriate safeguards for the rights and freedoms of third parties".[88]

However, because it is unclear who will enforce this provision, it raises the important question about who gets to decide which content falls under these categories. Divergences in the interpretation of these exceptions will likely allow many deep fakes to remain unlabelled.[89] Given the scale of distribution of online content, and the experience with the Code of Practice in labelling disinformation, it is probable that online platforms will play an important role in this task. Even more, since there is no clear sanction on non-compliance with the transparency obligations under Art. 52 (3), which is applicable to deep fakes, enforcement therefore will depend entirely on platforms commitment to abide by this proposed

86 *Michael Veale* and *Frederik Zuiderveen Borgesius*, 'Demystifying the Draft EU Artificial Intelligence Act' (SocArXiv, 5 July 2021) <https://osf.io/preprints/socarxiv/38p5f/>.

87 ibid 20.

88 Artificial Intelligence Act (n 13).

89 *European Parliamentary Research Service Scientific Foresight Unit* (STOA) (n 7).

rule. It is recommended that the final version of the AI Act will include clear incentives for compliance with transparency requirements when it comes to limited risk AI applications.

Moreover, and as a final point on enforcement, the AI Act does not consider complaint mechanisms for AI system providers, and it lacks a mechanism to hold regulators accountable for weak enforcement.[90] For example, users or human rights advocacy groups concerned about the potential harms of an AI system would not be able to request an investigation on products such as *DeepNude*,[91] that "undresses" photos of women, or other questionable deep fake systems that potentially compromises fundamental rights and that could be deployed in the EU market.[92]

b) Concerns over dual use of deep fakes in a risk-based assessment framework

Another challenging issue with the introduction of deep fakes in the AI Act is the notion of a high-risk system. An example of why laws regarding the production and use of deep fake technology are confusing is found in Annex III of the proposed AI Act, which lists high-risk AI systems referred to in Art. 6 (2). The provision includes that software used to detect deep fakes by law enforcement is included as high-risk raising the question of why the software that creates them is not. This is a particular question because current deep fake detection systems seem to spur deep fake creation systems. Deep fake detection is also recurrently compared in literature as an arms race due to the adversarial machine learning methods regarding deep fake creation systems. As indicated by *Greenwood*, it is true that these current assumptions may soon fail as the technology evolves and new models and processes for creating deep fakes are out that can strain or discontinue in some cases the adversarial relation between detection and creation.[93] It is also true, as explained by *Nguyen*, that a system used for creation can be separate to the one used for detection by relying on different technologies. In this case the law is regulating uses as opposed to the technology itself. Therefore, it is not to assume that the rules that reign over deep fake creation systems should fit detection systems since they can be fundamentally different. However, it is not in the scope of this paper to elucidate the technical specifications that differentiate both types of systems, only to question their subjective classification under the proposed AI Act.

90 *Veale* and *Borgesius* (n 86) 25.

91 *Karen Hao*, 'An AI App That "Undressed" Women Shows How Deepfakes Harm the Most Vulnerable' (2019) MIT Technology Review <https://www.technologyreview.com/2019/06/28/134352/an-ai-app-that-undressed-women-shows-how-deepfakes-harm-the-most-vulnerable/>.

92 *European Digital Rights* (EDRi), 'European Commission Adoption Consultation: Artificial Intelligence Act' (European Digital Rights 2021) 28.

93 *Daniel Greenwood* is a lecturer and research scientist at the MIT Media Lab with whom the author presented on deep fakes at a joined event between Legal Hackers Luxembourg, Moscow and Boston in December 2020. In the context of their exchanges this particular question was discussed.

Moreover, earlier drafts of the proposed AI Act indicated that an AI system could be considered high-risk if it had "systemic adverse impacts for society at large, including by endangering the functioning of democratic processes and institutions and the civic discourse, the environment, public health, [public security]."[94] As it has been discussed, one of the recognized harms as well as one of the risks of deep fakes is endangering the functioning of democratic processes, distorting democratic discourse, and compromising electoral integrity. Furthermore, the draft also contemplated that if there was an adverse impact on fundamental rights, then the AI system should be considered high-risk. This is precisely the case of fake non-consensual deep fakes, which are predominant and disproportionately harm women. Therefore, there seemed to be enough evidence under the first draft of the AI Act to qualify deep fake creation systems as high-risk. However, the final proposal deleted these criteria and replaced them with two conditions that must be met to be considered high-risk in Art. 6 (1). In its current form, none of the criteria could be met by deep fake creation systems. Nevertheless, the motivation to differentiate in terms of risk between deep fake detection systems as high-risk, while deep fake creation is not, needs to be clarified.

Also, on this point, the assumption that deep fake detection systems should be confined to law enforcement authorities seems flawed. Since media outlets and fact-checker organizations must be able to use these systems to debunk manipulated media and thus, contribute to fighting disinformation. First, if using a deepfake detection system imposes a high regulatory burden, the law potentially discourages its use, in particular because the AI Act proposal foresees stricter obligations on high-risk system providers. However, for example, fact-checker organizations or human rights advocacy groups might not be well-resourced to deal with complex requirements.[95] Second, providers of deep fake technology are potentially more adequate to carry out a disclosure obligation. Labels can be standardized and built-in into the systems and also "watermarking tools can also be integrated into devices that people use to make digital contents to create immutable metadata for storing originality details such as time and location of multimedia contents as well as their untampered attestment."[96] Because the AI Act proposal focuses on standardization for some categories of AI systems, these technical solutions do not seem unfeasible. Although, determining intent on a deep fake requires a human review that can take account of the context.

94 Leaked AI Act, first available in POLITICO Pro (link unavailable), and republished in *Lomas* (n 67).

95 See more on how the media verification ecosystem considers they should be equipped with tools and training to detect manipulated media, such as deep fakes in *The Partnership on AI,* 'Manipulated Media Detection Requires More Than Tools: Community Insights on What's Needed' (13 July 2020) <https://www.partnershiponai.org/manipulated-media-detection-requires-more-than-tools-community-insights-on-whats-needed/>.

96 *Nguyen* and others (n 30) 8.

IV. Criminalizing malicious deep fakes

1. Deep fakes and revenge porn legislations

If one considers that 96% of deep fakes are fake non-consensual porn, then labelling does not seem a proportionate response by the legislator. As explained by *Franks* and *Waldman* there is no effective or convincing way to "speak back to a fraudulent representation that is virtually indistinguishable from a real depiction of an individual engaged in graphic intimate activity". They compare it to the "unauthorized publication of a person's actual nude image, the dissemination of a home address, or the disclosure of one's sexual orientation" and therefore conclude that deep fakes are not ideas that can simply be countered with different and better ideas.[97] Considering the difficulty in redressing the harm, one should consider whether minimal transparency requirements for deep fakes on the AI Act will be helpful to counter the majority of deep fakes circulating on the Internet. A label will not undo the damage caused to a victim of a non-consensual deep fake nor deter its creation.

Moreover, since the AI Act proposal has a horizontal approach, the disclosure requirement in the act needs to be coherent with national legislative requirements that might offer an effective remedy to victims of deep fakes, so there are no gaps within the laws. Regularly policymakers and scholars assimilate fake non-consensual deep fakes to revenge pornography (also known as revenge porn). But revenge pornography does not benefit from a harmonized approach in the EU, and only some Member States have adopted laws that target this behavior. Most importantly, from a substantive law point of view, revenge pornography is a different category from deep fakes, and laws regulating them could apply to some cases but will not adequately regulate the vast majority of situations.

Revenge pornography is a misnomer since it is not pornography *per se*, because the videos and images are not typically produced for pornographic purposes. In particular in cases of revenge pornography, images or videos are originally obtained with the consent of the person depicted. The intention behind publication and redistribution is primarily to cause distress and harassment to the victim. Therefore, the applicable laws align more closely to anti-harassment, privacy, or even copyright law rather than legislation concerning pornography.[98] Moreover, even if deep fakes or revenge pornography were to be covered in the pornography legal framework, this might be ineffective. Regulating adult pornography on the Internet has been one of the major challenges of applying the law to online environments. Adult pornography has largely remained unregulated globally on the Internet – primarily because of the difficulties of transposing offline laws to cyberspace, in particular when those laws rely, in principle, on national notions of obscenity and indecency.[99]

97 *Franks* and *Waldman* (n 9).

98 *Abhilash Nair*, 'The Regulation of Internet Pornography: Issues and Challenges' (1st edn, Routledge 2019) 201.

99 ibid 211.

Furthermore, legislators tend to overly rely on revenge pornography as the closest analogy to deep fakes to think of regulatory measures. As highlighted by *McGlynn* et al., the "predominant focus on 'revenge porn' skews the legislative provisions, most obviously by focusing on the motive of the perpetrator – the 'revenge' – and on non-consensual distribution only."[100] Even though, as shown above, revenge pornography and deep fakes are two different phenomena. Revenge pornography legislation does not account for the fake or deceptive elements inherent to deep fakes. More research is needed in this area since it is hard to ascertain without empirical evidence whether the appropriate legal response is in criminal or civil law or even in a completely different regulatory strategy.

As technology evolves, there have been calls for a more robust regulatory response to deep fakes that go beyond the conceptual framework of pornography or revenge pornography. For some scholars, such as *Edwards*, the law should not underestimate the threat of revenge pornography because of the permanence of those videos on the Internet.[101] The same logic is applicable to the case of deep fakes. *Edward* argues that the "right to be forgotten", derived from the Court of Justice of the European Union (CJEU) *Google v Spain* case,[102] could be invoked.[103] Treating this problem from the angle of data protection could be more effective for the non-consensual type of deep fakes, particularly since victims seek remedy rather than prosecution in court out of fear of exposing themselves. Moreover, in some cases, deep fake production implies that consent must be given by the persons being portrayed for specific disclosures and, in particular, if it means processing sensitive personal data, which is afforded a high level of protection in the EU under the GDPR.[104]

Additionally, deep fakes erode sexual privacy, leaving a lasting and distressing legacy for the victim, and this aspect deserves more attention.[105] For this reason, *Citron* proposes in the US an "uniform approach to sexual privacy that includes federal and state penalties

100 *Clare McGlynn* and *Erika Rackley*, 'Image-Based Sexual Abuse' (2017) 37 Oxford Journal of Legal Studies 534, 22.

101 *Lilian Edwards*, 'Revenge Porn: Why the Right to Be Forgotten Is the Right Remedy' (The Guardian, 29 July 2014) <http://www.theguardian.com/technology/2014/jul/29/revenge-porn-right-to-be-forgotten-house-of-lords>.

102 In that case the CJEU held that operators of search engines can be required to remove personal information from search results published by third party websites in order to protect privacy and protect the personal data of the data subject. However, the data subject's right to make that request must be balanced against the interest of the general public to access his or her personal information. See *Google Spain SL v Agencia Española de Protección de Datos* [2014] Court of Justice of the European Union C-131/12.

103 *Edwards* (n 101). See more generally *Information Law and Policy Center University of London*, 'Turing Lecture: Regulating Unreality (Deepfakes, Revenge-Pornography & Fake News) – Professor Lilian Edwards' <https://infolawcentre.blogs.sas.ac.uk/2019/07/23/turing-lecture-regulating-unreality-deepfakes-revenge-pornography-fake-news-professor-lilian-edwards/>.

104 General Data Protection Regulation (n 24)

105 "People's nude images are posted online without permission. Machine-learning technology is used to create digitally manipulated 'deep fake' sex videos that swap people's faces into pornography. At the heart of these abuses is an invasion of sexual privacy – the behaviors and expecta-

for privacy invaders, removes the statutory immunity from liability for certain content platforms, and works in tandem with hate crime laws."[106] She is not the first scholar to allude to the necessity for new specific laws or amending existing laws to respond the unique challenges presented by the internet transnational nature, in particular when it comes to pornographic content. In the US, "pornography laws generally can only assist where images are obscene in nature, which is a high threshold to meet".[107]

Due to the lack of laws to deal with revenge pornography and deep fakes, *McGlynn* et al. call for broader conceptual recognition of *image-based sexual abuse*. "This term covers all forms of taking, making and sharing nude or sexual images without consent, including threats to share and alter images."[108] Therefore, the term encompasses the recognition of harms beyond revenge pornography or non-consensual pornography, thus reflecting the harms of deep fakes better.[109]

So far, and as discussed in the previous sections, there is a lack of applicable laws that can effectively offer a remedy to victims of deep fakes. Victims are often left with a patchwork of provisions from IP law to data protection that can potentially offer relief depending on the case's specifics, but certainly not to the majority of victims. By only introducing a disclosure obligation, the AI Act proposal attempts to regulate only disinformation deep fakes without addressing the more popular type of deep fakes. At the very least, this should be clarified by the legislator in two regards.

First, because it is unclear that this type of transparency obligation will serve as a deterrent for the creations of deep fakes or the use of deep fake technology, as it has already been discussed. Therefore, the move of certain jurisdictions, such as the US and UK, towards a criminalization of deep fakes is understandable, even if it is considered a more draconian measure. The way in which this transparency measure will be articulated with a protendency for criminalizing deep fakes needs to be addressed by the legislator to maintain coherence in EU law.[110]

tions that manage access to, and information about, the human body; gender identity and sexuality; intimate activities; and personal choices. More often, women and marginalized communities shoulder the abuse. Sexual privacy is a distinct privacy interest that warrants recognition and protection." *Citron*, 'Sexual Privacy' (n 10). For more on the harms of deep fakes videos see *Clare McGlynn* and *Nicola Gavey*, 'Shattering Lives and Myths', 24.

106 *Citron*, 'Sexual Privacy' (n 10) 2.

107 *Nair* (n 98) 204.

108 *Clare McGlynn* and *Erika Rackley*, 'Policy Briefing on Law Commission Consultation on Intimate Image Abuse' (5 May 2021) <https://claremcglynn.files.wordpress.com/2021/05/mcglynnrackley-stakeholder-briefing-5-may-2021-final-1.pdf>.

109 *McGlynn* and *Rackley* (n 100).

110 Analyzing EU Member States national laws and proposals on revenge porn is out of the scope of this paper. However, some Member States have already criminalized revenge porn such as Malta, Italy or have expanded the scope of already existing legislation such as France and Germany in an effort to address it. Since revenge porn is used as analogy for non-consensual deep fakes, it is plausible that there is a tendency to criminalize non-consensual deep fakes.

Second, it is essential to clarify how platform liability is to be articulated in the context of deep fakes. On the one hand, most popular online platforms have taken public action to deal with deep fakes based on their content moderation policies that do not allow for nudity and, presumably, spot political and issue-based disinformation. Therefore, both strands of deep fake seem to be covered under their internal policies. On the other hand, the level of enforcement of these policies is not transparent and has been contested by public authorities, researchers, and users alike. There is a lack of public data, for example, on how many deep fakes circulate through these platforms or on how many are detected by automated filters. This lack of data on platform enforcement makes it hard to assess if platforms are taking appropriate responses to deep fake phenomena. Furthermore, most dedicated sites for sharing deep fakes are niche and located outside the scope of EU law.[111] In that context, it seems as if the proposal of the DSA, jointly with the new version of the Code of Practice on Disinformation, might provide some effective remedy to victims. The analysis of the DSA provisions is out of the scope of this article;[112] however, it is essential to avoid fragmentation between the obligations in the AI Act and the DSA.

2. Legal responses to deep fake phenomena: selected examples of non-EU jurisdictions

a) United States

The legal response in the US has mainly been to criminalize the use of malicious deep fakes. Some US States such as Virginia, Texas, and California have already taken measures. For instance, Virginia (Bill 2678) became one of the first places to cover deep fakes by expanding its non-consensual pornography ban to include deep fakes.[113] Unlike Virginia, Texas Senate passed Bill 751, which criminalizes deep fakes created "with intent to injure a candidate or influence the result of an election" and which are "published and distributed within 30 days of an election."[114] California's response was comprehensive since it passed

111 Even though major online platforms have committed to ban deep fakes, online discussion boards like *4chan*, *8chan* and *Voat* allow people to request and share deepfake porn. Moreover, a number of stand-alone sites have been created and they can receive around 20,000 unique visitors every day. See *Law Commission of England and Wales*, 'Intimate Image Abuse' (Law Com No 253, 2021).

112 For more on this see *Mark Cole, Christina Etteldorf* and *Carsten Ullrich*, 'Updating the Rules for Online Content Dissemination: Legislative Options of the European Union and the Digital Services Act Proposal' (1st edn, Nomos 2021).

113 *Adi Robertson*, 'Virginia's "Revenge Porn" Laws Now Officially Cover Deepfakes' (The Verge, 1 July 2019) <https://www.theverge.com/2019/7/1/20677800/virginia-revenge-porn-deepfakes-nonconsensual-photos-videos-ban-goes-into-effect>.

114 Tex. S.B. No 751 "Act relating to the creation of a criminal offense for fabricating a deceptive video with intent to influence the outcome of an election." (1 September 2019) <https://capitol.texas.gov/tlodocs/86R/billtext/html/SB00751E.htm>; *Kenneth Artz*, 'Texas Outlaws "Deepfakes" – but the Legal System May Not Be Able to Stop Them' (Texas Lawyer, 11 October

law AB 730 prohibiting deep fakes from influencing political campaigns and law AB 602, which addresses non-consensual pornography deep fakes.[115] These two distinct legislative acts in California aim at countering different socio-material harms of deep fakes. The range of different laws exposes that there is no holistic approach yet on regulating deep fakes.

At the federal level different initiatives have taken place. For example, The Identifying Outputs of Generative Adversarial Networks (IOGAN) Act[116] was enacted in December 2020 and aims at structuring public authorities' research and monitoring on deep fake technology development. In particular, it supports the development of standards related to deep fake technology and promotes partnerships with private sectors to support authenticity measures and enhance identification capabilities. Other federal level proposals that were discussed in regards to deep fakes in the past legislature include: The Deepfake Report Act of 2019[117], A Bill to Require the Secretary of Defense to Conduct a Study on Cyberexploitation of Members of the Armed Forces and Their Families and for Other Purposes[118] and The Defending Each and Every Person from False Appearances by Keeping Exploitation Subject (DEEP FAKES) to Accountability Act.[119] Those bills died or where vetoed during the legislative procedures, however some of their provisions might be included in other bills to come.

It is interesting to highlight that from those bills at the federal level only the Deepfakes Accountability Act,[120] proposed in June 2019, aimed to combat the spread of disinformation through restrictions on deep fake video alteration technology. It sought to place unauthorized digital recreations of people under the umbrella of unlawful impersonation

2019) <https://www.law.com/texaslawyer/2019/10/11/texas-outlaws-deepfakes-but-the-legal-system-may-not-be-able-to-stop-them/>.

115 *Davis Wright*, 'Two New California Laws Tackle Deepfake Videos in Politics and Porn' (DWT, 14 October 2019) <https://www.dwt.com/insights/2019/10/california-deepfakes-law>; California A.B. No 602 "Depiction of individual using digital or electronic technology: sexually explicit material: cause of action" (3 October 2019) <https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB602>; California A.B No 730 "Elections: deceptive audio or visual media" <https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730>.

116 *Anthony Gonzalez*, 'Identifying Outputs of Generative Adversarial Networks Act', H.R.4355 – 116th Congress (2019-2020) <https://www.congress.gov/bill/116th-congress/house-bill/4355>.

117 *Rob Portman*, 'Deepfake Report Act of 2019', S. 2065 – 116th Congress (2019-2020) <https://www.congress.gov/bill/116th-congress/senate-bill/2065>.

118 *Ben Sasse*, 'A Bill to Require the Secretary of Defense to Conduct a Study on Cyberexploitation of Members of the Armed Forces and Their Families, and for Other Purposes.', S. 1348 – 116th Congress (2019-2020) <https://www.congress.gov/bill/116th-congress/senate-bill/1348/text>.

119 *Yvette D Clarke*, 'Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019', H.R.3230 – 116th Congress (2019-2020) <https://www.congress.gov/bill/116th-congress/house-bill/3230/text>.

120 ibid.

statutes.[121] The reasoning was that if it was possible to identify the creator of a deep fake, then he can be prosecuted for a criminal offense. However, this bill was criticized because the Internet offers a high level of anonymity to deep fake creators. Therefore, practical enforcement of legal provisions targeting deep fake creators are considered weak and will not bring justice to the victims. Second, the act only aimed to criminalize deep fakes that are not disclosed and meet the requirement of intent. It did not deter the production of deep fakes, even those of non-consensual nature, if they were correctly labelled and thus, by fulfilling their labelling obligation, perpetrators could potentially avoid criminal liability. This latter resonates with the proposed AI Act transparency requirement since deep fakes are permitted as long as they add a disclosing element.

Furthermore, the proposed Deepfake Accountability Act also encouraged the preparation of an institutional setting to deal with deep fakes. For example, the US Attorney's Office would have been in charge of "false intimate depictions" to coordinate prosecution. It also created a task force at the Department of Homeland Security to monitor and cooperate with platforms. As already discussed, monitoring enforcement regarding transparency requirements under Title IV is still work in progress in the case of the proposed AI Act. Building institutional capacity and awareness on this issue is a crucial element that any legislative act considering regulating deep fakes should strongly promote. In this sense, the Deepfake Task Force Act,[122] which was introduced in August 2021, intends to establish a National Deepfake and Digital Provenance Task Force in the US. This bill is still under consideration at the US Congress and it reinforces the idea of an interagency coordination and cooperation to tackle deep fake phenomena. In particular, it proposes to adopt policies on content provenance and technology standards to reduce the proliferation and impact of deep fakes.

Again, the development of standards occupies a central place, signaling that this is potentially the new direction for US lawmakers regarding deep fake regulation in the short term. Moreover, it has also focused on developing media forensic capabilities to deal with the effects of deep fakes. This nascent trend is coherent with the overall direction of the EU AI Act proposal. Since the AI Act also mandates the development of standards for high-risk AI systems, which includes deep fakes systems intended to be used by law enforcement authorities, as indicated in Art. 6 (c) of Annex III of the AI Act.

121 *Devin Coldewey*, 'DEEPFAKES Accountability Act Would Impose Unenforceable Rules – but It's a Start' (TechCrunch, 13 June 2019) <https://techcrunch.com/2019/06/13/deepfakes-accountability-act-would-impose-unenforceable-rules-but-its-a-start/>.

122 *Rob Portman*, 'Deepfake Task Force Act', S. 2559 — 117th Congress (2021-2022) <https://www.congress.gov/bill/116th-congress/senate-bill/2559>.

b) United Kingdom

More closely related to the EU, in 2021, there was a petition for criminalizing manufacturing and distribution of deep fake pornography at the UK Parliament.[123] However, in 2018, the government tasked the Law Commission, a statutory independent body in charge of reviewing and recommending reforms to the law, to look into deep fakes in the context of pornography and review the existing criminal laws with respect to taking, making, and sharing intimate images without consent. The Law Commission acknowledges the lack of specific provisions, stating:

> "Currently, there is no single criminal offence in England and Wales that governs the taking, making and sharing of intimate images without consent. Instead, we have a patchwork of offences that have developed over time, most of which existed before the rise of the Internet and use of smartphones. Each offence has different definitions and fault requirements, and there are some behaviours that are left unaddressed. […] Inconsistency over what type of intimate images are covered. […] Sharing an altered image – usually involving adding someone's head to a pornographic image is also not covered."[124]

Among the aspects that the Law Commission is reviewing, are the creation and dissemination of "realistic intimate or sexual images to be created or combined with existing images",[125] in other words, deep fakes, under the existing criminal law. In the consultation paper of February 2021, the Law Commission proposes to broaden the scope of what is meant by intimate image abuse to include altered images such as sexualised photoshopping and deep fake pornography.[126] It also proposes to create four new offences in the criminal code. The consultation closed in May 2021. Therefore, the final recommendation for the reform is not yet available and is expected to be published in Spring 2022.

An important point from a criminal law perspective of this review by the Law Commission is that it should only deal with individual offenders since the UK Government has been actively working on online platform liabilities for dissemination of illegal and harmful content. Furthermore, the Law Commission is working concurrently to review the application of and potential reform to the communications offences under Section 1 of the Malicious Communications Act 1988 and Section 127 of the Communications Act 2003. Both reforms will potentially impact the regulatory landscape for deep fakes in the UK.

The fact that there are two main strands of deep fakes, one leading to disinformation and the other one to image-based sexual abuse, poses the question of whether there should be two legislation pieces in the EU targeting each type of use or if including them in horizontal legislation such as the AI Act could be an effective strategy. As seen in the US and

123 *Petitions – UK Government and Parliament,* 'Petition: Criminalise Manufacturing and Distributing Deep-Fake Pornography' <https://petition.parliament.uk/petitions/567793>.

124 *Law Commission* (n 111).

125 *Law Commission of England and Wales*, 'Reform of the Communications Offences' (Law Com No 399, 20 July 2021).

126 ibid.

UK examples, the trend indicates that several pieces of legislation are proposed for each strand of deep fakes, while at the same time making sure platform liability issues are dealt with separately. Nevertheless, the criminalization of image-based sexual abuse is on the rise.

V. Conclusion

High-profile deep fake cases have caught the media's attention over the last year, highlighting the different concerns over this AI-based technology. Overall the debate in the media concerning deep fakes oscillates between those portraying a dystopian future in which deep fakes will become ubiquitous and almost impossible to detect and skeptics. Even though a wave of political deep fakes has not materialized yet, the number of non-consensual deep fakes grows steadily year after year. It is the fast-paced trend on the commodification of deep fakes that alarms experts and regulators. However, there are several technical measures that can help counter deep fakes for organizations, though the same cannot be said of laws. As it was presented in this contribution, deep fakes inhabit a regulatory void, where on a case-by-case basis some provisions of data protection, intellectual property, defamation, revenge porn, and platform liability laws could apply but do not comprehensively offer victims redress or deter malicious uses of deep fakes. It is in that space that the proposed AI Act attempts to regulate, minimally, some aspects of this technology at the EU level.

The AI Act proposal raises important questions on what deep fakes are, where the line should be drawn between lower AV manipulation and AI-based manipulation, and if any line should be drawn at all, considering that lower AV manipulation can be as harmful to individuals and society as deep fakes. Developing a common understanding of what deep fakes are is important to provide a shared vocabulary for regulators, scholars, online platforms, and other stakeholders to discuss online disinformation and manipulation problems, which is much needed. Currently, the consensus of academics and industry stakeholders is that two elements determine deep fakes: first, they are based on AI techniques, more precisely on deep learning to automate some parts of the audiovisual editing process, and second, they intend to deceive. Real or not, the uncertainty deep fakes create has a cost on society and causes a distinct profile of harm to individuals who compromise their agency over themselves. This article aimed to clarify some of the main issues and challenges of building a legal definition that adequately encompasses the harms of deep fakes and not only focuses on the use and the risk profile of an AI system. One of the points the proposal needs to improve is defining the legitimate purposes and malicious uses of deep fakes. Legitimate uses of deep fakes should not be subject to an unnecessary regulatory burden.

Moreover, the proposed AI Act regards the production of deep fakes as a low-risk AI application. Even though this categorization is subjective and questionable, the obligations concerning deep fakes are minimal transparency requirements under Art. 52 (3) of the AI Act. This article argues that the transparency requirements of the provision are not effective enough to counter deep fakes creation, dissemination, and harms, as shown by the use of

similar disclosure and labelling actions implemented under the Code of Practice on Disinformation for political advertising and coronavirus-related disinformation. This evidence becomes most alarming in the context of deep fakes, since experts agree that deep fakes are more persuasive content than fake news. Therefore, more research is needed into the effectiveness of labelling and other transparency requirements for online platforms and users, and how these measures could be improved. The scope of this provision is also broad and vague, and more refinement is needed for its practical implementation.

Another serious concern on the obligations regarding deep fakes is enforcement. Overall, enforcement of the proposed AI Act falls upon bodies with limited experience in the field of disinformation or in handling sexual-image-based abuse, two of the predominant issues with deep fakes. Not to mention that users, not providers, are subject to this provision and that there seems not to be any sanctions for non-compliance with disclosure obligations, sentencing the whole provision to weak enforcement. In that sense, to enhance enforcement, a broader range of policy options that do not revolve around deep fake detection should be discussed. A strengthened Code of Practice on Disinformation could be the tool to impose deep fake monitoring and additional obligations to online platforms for this area.

Furthermore, this contribution showed a brief comparative insight of US and UK legal approaches to deep fake phenomena that focus initially on regulating non-consensual deep fakes. Their response has been to criminalize them along the lines of laws against revenge pornography, even though these are two different issues. More recently, both jurisdictions have focused on disinformation deep fakes that can compromise electoral integrity. This illustrates a trend on separating regulation for the two main strands of deep fakes: disinformation and non-consensual pornography. The AI Act does not necessarily clarify if it aims only to counter disinformation risks, although it seems to be that way. If that is the case, this needs to be explicit, and the legislators need to foresee additional measures to curtail non-consensual deep fakes. Legislators should consider the gendered dimension of deep fakes when drafting a regulation targeting this technology. Finally, deep fakes are challenging to regulate because they embody elements of the much larger multi-faceted problem of disinformation and cyber misogyny in society.

Bibliography

Artz K, 'Texas Outlaws 'Deepfakes'—but the Legal System May Not Be Able to Stop Them' (*Texas Lawyer*, 11 October 2019) <https://www.law.com/texaslawyer/2019/10/11/texas-outlaws-deepfakes-but-the-legal-system-may-not-be-able-to-stop-them/>

AWO Agency, 'Tackling Gender Based Online Violence in the Digital Services Act' (*AWO Agency Online Event*, 30 September 2021)

BBC News, 'Fake Voices "Help Cyber-Crooks Steal Cash"' (*BBC News*, 8 July 2019) <https://www.bbc.co.uk/news/technology-48908736>

Brady M, 'Deepfakes: A New Disinformation Threat' (*Democracy Reporting International*, 2020) <https://democracy-reporting.org/wp-content/uploads/2020/08/2020-09-01-DRI-deepfake-publication-no-1.pdf>

Burkell J and Gosse C, 'Nothing New Here: Emphasizing the Social and Cultural Context of Deepfakes' (2019) *First Monday 24*(12) <https://firstmonday.org/ojs/index.php/fm/article/view/10287>

California A.B. No 602 "*Depiction of individual using digital or electronic technology: sexually explicit material: cause of action*" (3 October 2019) <https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB602>

California A.B No 730 "*Elections: deceptive audio or visual media*" (3 October 2019) <https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=201920200AB730>

Chesney R, 'Facebook Takes a Step Forward on Deepfakes – And Stumbles' (*Lawfare*, 8 January 2020) <https://www.lawfareblog.com/facebook-takes-step-forward-deepfakes-and-stumbles>

Chesney R and Citron D, 'Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security' (2019) *107 California Law Review* 1753

Chowdhury A, 'Deepfakes – Press Release' (*Future Advocacy*, 12 November 2019) <https://futureadvocacy.com/deepfakes-press-release/>

Citron D, '*Sexual Privacy*' (Faculty Scholarship, 2018) <https://digitalcommons.law.umaryland.edu/fac_pubs/1600>

——, '*Hate Crimes in Cyberspace*' (Harvard University Press, 2014) <http://www.degruyter.com/document/doi/10.4159/harvard.9780674735613/html>

Citron D and Franks MA, 'Criminalizing Revenge Porn' (2014) *49 Wake Forestlaw Review* 48

Clarke YD, 'Defending Each and Every Person from False Appearances by Keeping Exploitation Subject to Accountability Act of 2019', *H.R.3230 – 116th Congress* (2019-2020),<https://www.congress.gov/bill/116th-congress/house-bill/3230/text>

Coldewey D, 'DEEPFAKES Accountability Act Would Impose Unenforceable Rules — but It's a Start' (*TechCrunch,* 13 June 2019) <https://techcrunch.com/2019/06/13/deepfakes-accountability-act-would-impose-unenforceable-rules-but-its-a-start/>

Cole M, Etteldorf C and Ullrich C, '*Cross-Border Dissemination of Online Content – Current and Possible Future Regulation of the Online Environment with a Focus on the EU E-Commerce Directive (Open Access)*' (Nomos, 2020)

Cole M, Etteldorf C and Ullrich C, '*Updating the Rules for Online Content Dissemination: Legislative Options of the European Union and the Digital Services Act Proposal*', vol 83 (1st edn, Nomos, 2021) <https://doi.org/10.5771/9783748925934>

Court of Justice of the European Union [2014] C-131/12 *Google Spain SL v Agencia Española de Protección de Datos*

Diakopoulos N and Johnson D, 'Anticipating and Addressing the Ethical Implications of Deepfakes in the Context of Elections' (2019) *New Media & Society* <https://papers.ssrn.com/abstract=3474183>

Dignum V and Muller C, '*Artificial Intelligence Act: Analysis and Recommendations*' (ALLAI 2021)

Dolhansky B and others, '*The DeepFake Detection Challenge (DFDC) Dataset*' (2020) arXiv:2006.07397 [cs] <http://arxiv.org/abs/2006.07397>

Edelman G, 'Facebook's Deepfake Ban Is a Solution to a Distant Problem' (*Wired,* 7 January 2020*)* <https://www.wired.com/story/facebook-deepfake-ban-disinformation/>

Edwards L, 'Revenge Porn: Why the Right to Be Forgotten Is the Right Remedy' (*The Guardian*, 29 July 2014) <http://www.theguardian.com/technology/2014/jul/29/revenge-porn-right-to-be-forgotten-house-of-lords>

ERGA, 'ERGA Report on Disinformation: Assessment of the Implementation of the Code of Practice' (*European Regulators Group for Audiovisual Media Services*, 2020)

——, 'ERGA Recommendations for the New Code of Practice on Disinformation' (*European Regulators Group for Audiovisual Media Services*, 2021)

EU Newsroom, '*2021 Vademecum of the Assembly of the Signatories of the Code of Practice on Disinformation*' <https://ec.europa.eu/newsroom/repository/document/2021-27/Vademecum_2021_Code_poQcw5VoJg362zUKq6VL774dsFs_78161.pdf>

European Commission, 'Commission Appoints Expert Group on AI and Launches the European AI Alliance' (*Shaping Europe's digital future*, 14 June 2018) <https://digital-strategy.ec.europa.eu/en/news/commission-appoints-expert-group-ai-and-launches-european-ai-alliance>

——, 'Tackling Online Disinformation: A European Approach' COM (2018) 236 final

——, 'White Paper "On Artificial Intelligence – A European Approach to Excellence and Trust"' COM (2020) 65 final

——, 'Impact Assessment Accompanying the Proposal for a Regulation of the European Parliament and of the Council Laying Down Harmonised Rules On Artificial Intelligence (Artificial Intelligence Act) And Amending Certain Union Legislative Acts' SWD (2021) 84 final

——, 'First Baseline Reports – Fighting COVID-19 Disinformation Monitoring Programme' (*Shaping Europe's digital future,* 10 September 2020) <https://digital-strategy.ec.europa.eu/en/library/first-baseline-reports-fighting-covid-19-disinformation-monitoring-programme>

——, 'Guidance to Strengthen Code of Practice on Disinformation' COM (2021) 262 final

European Digital Rights (EDRi), 'European Commission Adoption Consultation: Artificial Intelligence Act' (*European Digital Rights*, 2021)

European Parliament and the Council, 'Proposal for a Regulation of the European Parliament and of the Council Laying down harmonised rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain Union legislative acts' COM (2021) 206 final

——, 'Proposal for a Regulation of the European Parliament and of the Council on a Single Market For Digital Services (Digital Services Act) and amending Directive 2000/31/EC' COM (2020) 825 final

——, 'Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation)' (2016) OJ L 119/1.

European Parliamentary Research Service Scientific Foresight Unit (STOA), 'Tackling Deepfakes in European Policy' (*EU Publications Office*, 2021) PE 690.039

Facebook, '*Reports on Implementation of the Code of Practice on Disinformation*' (2019)

——, 'Enforcing Against Manipulated Media' (*About Facebook*, 7 January 2020) <https://about.fb.com/news/2020/01/enforcing-against-manipulated-media/>

Floridi L, 'Artificial Intelligence, Deepfakes and a Future of Ectypes' (2018) *31 Philosophy & Technology* 317

Foer F, 'The Era of Fake Video Begins' (*The Atlantic*, 8 April 2018) <https://www.theatlantic.com/magazine/archive/2018/05/realitys-end/556877/>

Foley J, '14 Deepfake Examples That Terrified and Amused the Internet' (*Creative Bloq*) <https://www.creativebloq.com/features/deepfake-examples>

Franks MA and Waldman A, 'Sex, Lies, and Videotape: Deep Fakes and Free Speech Delusions' (2019) *78 Md. L. Rev.* 892

Gold H, 'Facebook Tries to Curb Deepfake Videos as 2020 Election Heats Up' (*CNN*, 7 January 2020) <https://edition.cnn.com/2020/01/07/tech/facebook-deepfake-video-policy/index.html>

Goldman E, 'Content Moderation Remedies' (2021) *28 Michigan Technology Law Review* 1, Santa Clara Univ. Legal Studies Research Paper <https://papers.ssrn.com/abstract=3810580>

Gonzalez A, 'Identifying Outputs of Generative Adversarial Networks Act', *H.R.4355 – 116th Congress* (2019-2020) <https://www.congress.gov/bill/116th-congress/house-bill/4355>

Hao K, 'An AI App That "Undressed" Women Shows How Deepfakes Harm the Most Vulnerable' (2019) *MIT Technology Review* <https://www.technologyreview.com/2019/06/28/134352/an-ai-app-that-undressed-women-shows-how-deepfakes-harm-the-most-vulnerable/>

——, 'AI Voice Actors Sound More Human than Ever—and They're Ready to Hire' (2021) *MIT Technology Review* <https://www.technologyreview.com/2021/07/09/1028140/ai-voice-actors-sound-human/>

Hildebrandt M, 'A Commentary on the Proposal for an EU AI Act of 21 April 2021' (*Vrije Universiteit Brussel*, 19 July 2021)

Huck Magazine, 'How Carrie Goldberg Turned Litigation into an Act of Protest' (*Huck Magazine,* 20 August 2019) <https://www.huckmag.com/art-and-culture/books-art-and-culture/how-carrie-goldberg-turned-litigation-into-activism/>

Information Law and Policy Center University of London, '*Turing Lecture: Regulating Unreality (Deepfakes, Revenge-Pornography & Fake News) – Professor Lilian Edwards*' <https://infolawcentre.blogs.sas.ac.uk/2019/07/23/turing-lecture-regulating-unreality-deepfakes-revenge-pornography-fake-news-professor-lilian-edwards/>

Kaiser B, Mayer J and Matias JN, 'Warnings That Work: Combating Misinformation Without Deplatforming' (*Lawfare*, 23 July 2021) <https://www.lawfareblog.com/warnings-work-combating-misinformation-without-deplatforming>

Knight W, 'Facebook Is Making Its Own AI Deepfakes to Head off a Disinformation Disaster' (2019) *MIT Technology Review* <https://www.technologyreview.com/2019/09/05/65353/facebook-is-making-ai-deepfakes-to-head-off-a-disinformation-disaster/>

KrASIA, 'Did Myanmar's Military Deepfake a Minister's Corruption Confession?' (*KrASIA*, 24 March 2021) <https://kr-asia.com/did-myanmars-military-deepfake-a-ministers-corruption-confession>

Kumar R, Sotelo J, Kumar K, De Brébisson A and Bengio Y, Jos, 'ObamaNet: Photo-Realistic Lip-Sync from Text' (*NeurIPS*, 2017) <https://ritheshkumar.com/obamanet/>

Law Commission of England and Wales, 'Intimate Image Abuse' (*Law Com No 253*, 2021)

——, 'Reform of the Communications Offences' (*Law Com No 399*, 20 July 2021) <https://www.lawcom.gov.uk/project/reform-of-the-communications-offences/>

Lomas N, 'EU Plan for Risk-Based AI Rules to Set Fines as High as 4% of Global Turnover, per Leaked Draft' (*TechCrunch,*14 April 2021) <https://techcrunch.com/2021/04/14/eu-plan-for-risk-based-ai-rules-to-set-fines-as-high-as-4-of-global-turnover-per-leaked-draft/>

——, 'Facebook Bans Deceptive Deepfakes and Some Misleadingly Modified Media' (*TechCrunch*, 7 January 2020) <https://social.techcrunch.com/2020/01/07/facebook-bans-deceptive-deepfakes-and-some-misleadingly-modified-media/>

The Partnership on AI, 'Manipulated Media Detection Requires More Than Tools: Community Insights on What's Needed' (*The Partnership on AI,* 13 July 2020) <https://www.partnershiponai.org/manipulated-media-detection-requires-more-than-tools-community-insights-on-whats-needed/>

Marsden C, Meyer T and Brown I, 'Platform Values and Democratic Elections: How Can the Law Regulate Digital Disinformation?' (2020) *36 Computer Law & Security Review* 105373

McGlynn C and Rackley E, 'Image-Based Sexual Abuse' (2017) *37 Oxford Journal of Legal Studies* 534

——, '*Policy Briefing on Law Commission Consultation on Intimate Image Abuse*' (5 May 2021) <https://claremcglynn.files.wordpress.com/2021/05/mcglynnrackley-stakeholder-briefing-5-may-2021-final-1.pdf>

McGlynn PC and Gavey PN, '*Shattering Lives and Myths: A Report on Image-Based Sexual Abuse*' (1 July 2019) <https://claremcglynn.files.wordpress.com/2019/10/shattering-lives-and-myths-revised-aug-2019.pdf>

Meta AI, 'Deepfake Detection Challenge Results: An open initiative to advance AI' (*Meta AI*, 12 June 2020) <https://ai.facebook.com/blog/deepfake-detection-challenge-results-an-open-initiative-to-advance-ai/>

Nair A, '*The Regulation of Internet Pornography: Issues and Challenges*' (1st edn, Routledge, 2019)

Nguyen TT and others, '*Deep Learning for Deepfakes Creation and Detection: A Survey*' (2021) arXiv:1909.11573 [cs, eess] <http://arxiv.org/abs/1909.11573>

Paris B and Donovan J, 'Deepfakes and Cheap Fakes: The Manipulation of Audio and Visual Evidence' (*Data & Society*, 2019)

Patrini G, '*Automating Image Abuse: Deepfake Bots on Telegram*' (20 October 2020) <https://giorgiop.github.io/posts/2020/10/20/automating-image-abuse/>

Patrini G, Cavalli F and Adjer H, 'The State of Deepfakes: Reality under Attack' (*Deeptrace*, 2018) Annual Report v2.3

Perot E and Mostert F, 'Fake It Till You Make It: An Examination of the US and English Approaches to Persona Protection As Applied to Deepfakes on Social Media' (*SSRN*, 2020) ID 3537052 <https://papers.ssrn.com/abstract=3537052>

Petitions – UK Government and Parliament, '*Petition: Criminalise Manufacturing and Distributing Deep-Fake Pornography*' <https://petition.parliament.uk/petitions/567793>

Portman R, 'Deepfake Report Act of 2019', S. 2065 – *116th Congress* (2019-2020) <https://www.congress.gov/bill/116th-congress/senate-bill/2065>

Portman R, 'Deepfake Task Force Act', S. 2559 – *117th Congress* (2021-2022) <https://www.congress.gov/bill/116th-congress/senate-bill/2559>

Resnick B, 'We're Underestimating the Mind-Warping Potential of Fake Video' (*Vox*, 20 April 2018) <https://www.vox.com/science-and-health/2018/4/20/17109764/deepfake-ai-false-memory-psychology-mandela-effect>

Reuters, 'Fact Check: Donald Trump Concession Video Not a "Confirmed Deepfake"' (*Reuters*, 11 January 2021) <https://www.reuters.com/article/uk-factcheck-trump-consession-video-deep-idUSKBN29G2NL>

Robertson A, 'Virginia's "Revenge Porn" Laws Now Officially Cover Deepfakes' (*The Verge*, 1 July 2019) <https://www.theverge.com/2019/7/1/20677800/virginia-revenge-porn-deepfakes-nonconsensual-photos-videos-ban-goes-into-effect>

Ruiter A, 'The Distinct Wrong of Deepfakes' (2021) *Philosophy & Technology*

Sasse B, 'A Bill to Require the Secretary of Defense to Conduct a Study on Cyberexploitation of Members of the Armed Forces and Their Families, and for Other Purposes.', S. 1348 – *116th Congress* (2019-2020) <https://www.congress.gov/bill/116th-congress/senate-bill/1348/text>

Sensity AI, 'The State of Deepfakes 2019. Landscape, Threats, and Impact' (*Sensity AI*, 2019) <https://sensity.ai/reports/>

Somers M, 'Deepfakes, Explained' (*MIT Sloan*, 21 July 2020) <https://mitsloan.mit.edu/ideas-made-to-matter/deepfakes-explained>

Synced, 'NVIDIA Open-Sources Hyper-Realistic Face Generator StyleGAN' (*SyncedReview*, 9 February 2019) <https://medium.com/syncedreview/nvidia-open-sources-hyper-realistic-face-generator-stylegan-f346e1a73826>

Tex. S.B. No 751, "Act relating to the creation of a criminal offense for fabricating a deceptive video with intent to influence the outcome of an election" (1 September 2019) <https://capitol.texas.gov/tlodocs/86R/billtext/html/SB00751E.htm>

The Guardian, 'Facebook Bans "deepfake" Videos in Run-up to US Election' (*The Guardian*, 7 January 2020) <http://www.theguardian.com/technology/2020/jan/07/facebook-bans-deepfake-videos-in-run-up-to-us-election>

The Guardian, 'Facebook Refuses to Remove Doctored Nancy Pelosi Video' (*The Guardian*, 3 August 2020) <http://www.theguardian.com/us-news/2020/aug/03/facebook-fake-nancy-pelosi-video-false-label>

The Takeaway, 'Are Deepfakes the Next Fake News?' (*WNYC Studios*, 22 July 2019) <https://www.wnycstudios.org/podcasts/takeaway/segments/deepfakes-fake-news-artificial-intelligence>

Twitter, 'Our Synthetic and Manipulated Media Policy' (*Twitter Help*) <https://help.twitter.com/en/rules-and-policies/manipulated-media>

Ullrich C, '*Unlawful Content Online: Towards A New Regulatory Framework For Online Platforms*', vol 21 (Nomos, 2021)

United Nations Interregional Crime and Justice Research Institute (UNICRI), 'Algorithms and Terrorism: The Malicious Use of Artificial Intelligence for Terrorist Purposes' (*United Nations Office of Counter-Terrorism* (UNOCT), 2021) <http://unicri.it/News/Algorithms-Terrorism-Malicious-Use-Artificial-Intelligence-Terrorist-Purposes>

Veale M and Borgesius FZ, 'Demystifying the Draft EU Artificial Intelligence Act' (*SocArXiv*, 5 July 2021) <https://osf.io/preprints/socarxiv/38p5f/>

Vincent J, 'Why We Need a Better Definition of "Deepfake"' (*The Verge*, 2018) <https://www.theverge.com/2018/5/22/17380306/deepfake-definition-ai-manipulation-fake-news>

——, 'A Spy Reportedly Used an AI-Generated Profile Picture to Connect with Sources on LinkedIn' (*The Verge*, 13 June 2019) <https://www.theverge.com/2019/6/13/18677341/ai-generated-fake-faces-spy-linked-in-contacts-associated-press>

Wardle C and Derakhshan H, Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making (*Council of Europe*, 2017) DGI(2017)09

WITNESS, 'Prepare, Don't Panic: Synthetic Media and Deepfakes' (*WITNESS Media Lab*) <https://lab.witness.org/projects/synthetic-media-and-deep-fakes/>

Wright D, 'Two New California Laws Tackle Deepfake Videos in Politics and Porn' (*DWT*, 14 October 2019) <https://www.dwt.com/insights/2019/10/california-deepfakes-law>