ACIG
APPLIED CYBERSECURITY & INTERNET GOVERNANCE

NASK

Regulating Deep Fakes in the Artificial Intelligence Act

Mateusz Łabuz | Ministry of Foreign Affairs of the Republic of Poland, Chemnitz University of Technology, Germany, ORCID: 0000-0002-6065-2188

Abstract

The Artificial Intelligence Act (AI Act) may be a milestone in the regulation of artificial intelligence by the European Union. The regulatory framework proposed by the European Commission has the potential to serve as a global benchmark and strengthen the position of the EU as one of the main players on the technology market. One of the components of the draft regulation are the provisions on deep fakes, which include a relevant definition, risk category classification and transparency obligations. Deep fakes rightly arouse controversy and are a complex phenomenon. When leveraged for negative purposes, they significantly increase the risk of political manipulation, and at the same time contribute to disinformation, undermining trust in information and the media. The AI Act may strengthen the protection of citizens against some of the negative consequences of misusing deep fakes, although the impact of the regulatory framework in its current form will be limited due to the specificity of their creation and dissemination. The effectiveness of the provisions will depend not only on enforcement capabilities, but also on the precision of phrasing provisions to prevent misinterpretation and deliberate abuse of exceptions. At the same time, the AI Act will not cover a significant portion of deep fakes, which, due to the malicious intentions of their creators, will not be subject to the transparency obligations. This study analyses provisions related to deep fakes in the AI Act and proposes improvements that will take into account the specificity of this phenomenon to a greater extent.

Keywords

deep fakes, Artificial Intelligence Act, AI Act, regulations, European Union, transparency obligations, disclosure rules

Received: 14.09.2023

Accepted: 19.10.2023

Published: 27.10.2023

Cite this article as:
M. Łabuz "Regulating Deep Fakes in the Artificial Intelligence Act," ACIG, vol. 2, no. 1, 2023, DOI: 10.60097/ACIG/162856

Corresponding author:
Mateusz Łabuz, Ministry of Foreign Affairs of the Republic of Poland, Chemnitz University of Technology, Germany
ORCID: 0000-0002-6065-2188; E-MAIL: mateusz.labuz@msz.gov.pl

ACIG
APPLIED CYBERSECURITY & INTERNET GOVERNANCE

1. Introduction

The proposal for Regulation of the European Parliament and the Council laying down harmonised rules on Artificial Intelligence (AI Act)[1], introduced by the European Commission in April 2021, is intended to be one of the key elements in positioning the European Union to regulate the dynamic development of artificial intelligence (AI). Creating a legal framework for AI will not only allow the EU to face numerous legal, political, economic and social challenges, but will also put it in a privileged position in the global competition to set regulatory standards [1, 2], and perhaps even to 'serve as a benchmark for other countries' [3]. The AI Act addresses the risks associated with certain uses of technology and aims to achieve 'the development of an ecosystem of trust by proposing a legal framework for trustworthy AI' [4]. Creating 'an ecosystem of trust' is an ambitious task, which requires internal consistency and a great sense in using legal terms, so they are not contested or interpreted in a way that is incongruent with the spirit of the regulation.

1 — As of the date this paper was written (August 2023), the AI Act was still being negotiated.

Deep fakes, which first began to appear in 2017, are a relatively well-described phenomenon and have been the subject of numerous analyses that, among others, extensively described the various ways they are used to inflict harm [5–8]. The potential risks of the misuse of deep fakes include the spread of fake news and disinformation, election manipulation, creation of non-consensual pornographic content, defamation, discredit and ridicule of individuals, including political opponents, undermining trust in traditional media messages, distortion of reality, impairment of political engagement within society, undermining of the epistemic quality of debate and thus democratic discourse, threats to the stability of economic systems, spread of hate speech and strengthening gender inequalities, as well as psychological harm to individuals or vulnerable groups [9–11].

This non-exhaustive enumeration does not fully reflect the specificity of the phenomenon. One should not forget that deep fakes find many positive applications in the media, education, leisure and healthcare [12]. Therefore, they should not be demonised wholesale and every legal solution should take into account the diversity of uses and consequences related to the creation and dissemination of deep fakes in their various forms [13].

To date, deep fakes have somehow eluded basic legislation and rules governing their use have mostly been taken from provisions of civil, tort, criminal or copyright law [14]. The first attempts to regulate deep fakes in more specific legal acts should be observed with

interest, especially in terms of their implementation and impact on social and political processes. The EU can play a constructive role in this process, not only by referring to deep fakes in the AI Act, but also by using that leverage to introduce stricter countermeasures. The development of technology and frequently reported misuses require deep fakes to be directly regulated and, if necessary, forbidden if they directly violate the rights of third parties [16]. However, one should be realistic – even outright bans would not be completely effective since most deep fakes are meant to deceive recipients and circumvent legal, technical and social safeguards [17]. It is also necessary to consider how the law can protect the basic values of democracy from malicious and non-malicious uses of deep fakes, while preserving fundamental rights, including freedom of speech [18].

The AI Act refers to deep fakes explicitly, introducing a definition of the term, basic transparency and disclosure rules, and assigns deep fakes to the 'specific' or 'limited risk' quasi-category of AI systems [19, 20]. Some of the proposals introduced in 2021 by the Commission were rightly criticised by experts, who pointed to an insufficient legal regime, or underestimation of the seriousness of threats stemming from the creation and dissemination of deep fakes [21–23].

The amendments proposed by the European Parliament in June 2023 [24] have the potential to eliminate some of the deficits in the Commission's original proposal and can be generally deemed as a step in the right direction. At the same time, there are still shortcomings that should be addressed as part of further negotiations in order to create a coherent, although quite general in nature, legal framework for regulating deep fakes. However, it seems crucial to verify whether the proposed solutions will create an effective framework for combating deep fakes, which, in light of previous cases, seems doubtful and may force the EU to quickly examine and adjust its approach to their regulation.

The EU must ensure internal consistency, so that the definitions and solutions proposed in various documents are complementary and do not lead to misinterpretation or discrepancies. At the same time, deep fakes, due to their complexity and the cascading effects of their misuse [25], are a phenomenon that must be taken into account in more specific acts, which paves the way to further discussion on enforcement, liability and penalisation [26].

This study primarily serves to highlight the issue of deep fakes in light of the AI Act and is part of the current debate [9, 27–30] on the risks associated with the dissemination of technology that enables the

creation of hyper-realistic but fake synthetic media [10], which are increasingly difficult to distinguish from real ones [31, 32].

The aim of the study is to assess the current state of the AI Act in regard to deep fakes, as well as to draw attention to the shortcomings of the proposal. As already mentioned, deep fakes cannot be categorised unequivocally due to their multitude of applications. At the same time, it should be emphasised that they play an increasingly important role in entrenching digital disinformation [33, 34] and negatively affect many spheres of life [6, 9]. In some cases, they might directly threaten democracy, free elections and the information ecosystem, undermine trust in the media, or lead to the victimisation of individuals, especially women [9, 35]. The comments made by the author may also serve as a signpost for policymakers who, regardless of EU regulations, sooner or later will have to face the problem of deep fakes at the level of national legislation.

As the AI Act is still being negotiated, further changes to the substance of the regulation are possible, which might make it possible to eliminate deficits or shortcomings in already implemented measures.[2] Due to possible changes in the regulation, this study might become obsolete when the AI Act is adopted, which is a significant limitation. Nevertheless, analysis of the draft proposals and criticism of selected solutions can provide additional input for discussions on creating a regulatory framework in relation to deep fakes, which increases the paper's topicality and applicability.

2 —— The analysis was based on the proposals from April 2021 (the European Commission's proposal) and June 2023 (European Parliament amendments).

2. Definition scope

A holistic approach to the issue of deep fakes requires, first of all, the introduction of a legal definition of this term. The AI Act may be a point of reference for further work and legislation in this regard, which makes the EU's approach to the definition of key importance.

The Commission's proposal [4] referred only to a quasi-definition of 'deep fake'. Although the most relevant Article 3 contained definitions of terms used by the AI Act, deep fakes were not included in the list. The description of a deep fake was inserted into Article 52(3), which was supposed to set out transparency obligations for certain 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.*

Article 52(3) was later amended by the European Parliament to expand the range of the quasi-definition and introduce stricter transparency obligations:

> *Users of an AI system that generates or manipulates text, audio or visual content that would falsely appear to be authentic or truthful and which features depictions of people appearing to say or do things they did not say or do, without their consent ('deep fake'), shall disclose in an appropriate, timely, clear and visible manner that the content has been artificially generated or manipulated, as well as, whenever possible, the name of the natural or legal person that generated or manipulated it.*

The scope of this provision will be discussed at a later stage because the key change in the European Parliament's amendments in regard to defining deep fakes is the addition of a new point 44d in Article 3(1), which introduces the legal definition of the term and should be treated as a point of reference:

> *'deep fake' means manipulated or synthetic audio, image or video content that would falsely appear to be authentic or truthful, and which features depictions of persons appearing to say or do things they did not say or do, produced using AI techniques, including machine learning and deep learning.*

The addition of point 44d in Article 3(1) allows us to extract the four most important aspects of the definition of deep fakes:

1. technical, relating to the method of creation (manipulated or synthetic content, produced using AI techniques, including machine learning and deep learning);
2. typological, relating to the form of media that was used (audio, image or video content);
3. subjective, referring to the subject/object of depiction (features persons);
4. effectual, relating to the manner and effect of depiction (falsely appears to be authentic or truthful; appearing to say or do things they did not say or do).

Only meeting all aspect criteria together constitutes a deep fake. The survey conducted by A. Fernandez [21] to establish the elements of commonly used definitions of deep fakes resulted in the recognition of two mandatory features agreed on by scholars: 1) intervention by AI (which overlaps with the technical aspect); 2) the potential to deceive (which overlaps with the effectual aspect). Referring to A. de Ruiter [27], Fernandez considered the deceptive effect as a 'by-product of the creator's intent'. This approach seems only partially correct, as the very nature of a deep fake is based on the presentation of a false or distorted reflection of reality and thus implies intent to deceive recipients.

In general the decision to introduce a legal definition of 'deep fake' within the AI Act should be assessed positively. The Parliament's deletion of the wording 'to a person' is an unequivocally positive development in comparison to the Commission's phrasing of Article 52(3), because it expands the range of entities that can be targeted by audio or visual forgery, specifies the technical aspects, and makes explicit mention of machine learning and deep learning technologies. One could rightly argue that confirming whether content constituted a deep fake would require proving that an AI system was used to generate it. The degree of technological advancement with respect to tools used to create deep fakes is so high that unambiguous evidence that AI was used to generate content may be difficult or even impossible to obtain [21]. However, the definition does not seem to offer a reasonable alternative for the technical aspect and only the practical functioning of the provisions will reveal whether the classification of materials as deep fakes is rendered impossible by an inability to prove the use of AI.

To this extent, the definition in point 44d of Article 3(1) extends the scope of the quasi-definition included in the original Article 52(3) proposed by the Commission.

Intervention in the subjective aspect, which narrows the scope of subjects/objects depicted [24] with the wording *which features depictions of persons appearing to say or do things they did not say or do* is a negative development. Reference to objects, places or other entities or events that appeared in the original Commission proposal have been erroneously deleted, which limits the possibility of classifying content as a deep fake.

Meanwhile, there are deep fakes that do not depict people, but have proven to be effective tools in significantly influencing reality. In May 2023, an image deep fake depicting an explosion near the

Pentagon was disseminated via social media, leading to short-term losses on the New York Stock Exchange. According to Bloomberg, it was 'possibly the first instance of an AI-generated image moving the market' [36]. It is also possible to imagine images of natural disasters, military equipment, war damage, or desecration of religious symbols that do not include people [37]. Each one could serve as an inflammatory spark, leading to social unrest, or mobilisation of specific groups, and contribute to disinformation [38]. Some researchers are already warning of 'deep fake geography', which refers to falsification of cartographical data, including satellite images [39, 40]; some states have allegedly already used such images for the purpose of sophisticated disinformation [41].

Although some scenarios are for now only the subject of speculation, they are already being discussed by researchers, who are trying to raise awareness among policymakers. For this reason, extending the subjective scope of the definition of 'deep fake' and using the earlier proposal of the Commission seems advisable. Interestingly, the European Parliament's Committee on Industry, Research and Energy, in its opinion of June 2022 [42], proposed a deep fake definition that referred to *material that gives an authentic impression, in which events appear to be taking place, which never happened,* completely omitting any remarks concerning 'persons'.

In Recital 70 of the initial draft of the AI Act [4], which has not been amended by the European Parliament, the Commission referred

> *to certain AI systems intended to interact with natural persons or to generate content [that] may pose specific risks of impersonation or deception irrespective of whether they qualify as high-risk or not. And later: 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.*

Though deep fakes were not explicitly mentioned in Recital 70, they definitely match the description, which again features an extended subjective scope.

Unfortunately, there are significant differences between Recital 70, Article 3(1) point 44d and Article 52(3). These discrepancies need to be clarified in future.

Article 52(3) [24] leads to even more confusion as to whether the typological aspect was extended by the European Parliament to include text forms. Although deep fakes have no agreed-upon technical or typological definition [43], some concepts are circulating among scholars. While in the majority of the analysed studies, including reports from the European Union Agency for Cybersecurity (ENISA), EUROPOL, NATO, and an AI glossary by Brookings, the definitions are narrowed down and explicitly mention audio, image or video content [34, 44–47], some researchers also mention deep fakes in text form [28, 48–50], or are even concentrating on developing deep fake text detection methods [51].

At this stage, the AI Act does not completely resolve the problem of qualifying textual deep fakes. The European Parliament decided not to include text form of deep fakes within the scope of the definition included in Article 3(1). On the other hand, the amended Article 52(3), referring to transparency obligations, creates ambiguity, as the Commission's proposal was supplemented with the term 'text'. In the further part of the provision it was indicated that this might also refer to deep fakes. The literal understanding of the provision suggests that the scope of definition contained in Article 3(1) has been extended with respect to the typological aspect. Undoubtedly, appropriate disclosure rules should also apply to AI-generated or AI-manipulated texts, but they do not necessarily have to be qualified as deep fakes. This ambiguity needs to be clarified in future.

It is worth mentioning that the issue of extending the definition to include text deep fakes was raised by the European Parliament's Committee on Culture and Education [52], whose proposal for amendments of June 2022 referred to a deep fake as: *manipulated or synthetic audio, visual or audiovisual content, text or scripts which feature persons purported to be authentic and truthful.* A similar position (advocating for inclusion of deep fakes in text form) was taken by Mesarčík et al. [22] in a critical analysis of the AI Act proposal, though this study did not contain a proper rationale for such an inclusion.

In the author's opinion, the EU needs to either clearly include text deep fakes in the definition in Article 3(1), or clearly distinguish the text form from deep fakes, focusing only on audio and visual content in Article 52(3), and possibly create an additional provision for AI-generated and AI-manipulated texts. The latter seems to be the solution that would better match the analyses carried out by the majority of researchers.

Another problem in defining deep fakes in EU legal acts is the consistency of the proposed solutions. Deep fakes rarely receive an explicit

legal definition; more often regulations can be derived from specific formulations relating to phenomena that are similar or identical to deep fakes. The AI Act can serve as a benchmark for other legal acts, which in turn requires the consistent use of one term and one qualification. If the AI Act introduces the legal definition of deep fakes, other definitions or quasi-definitions/descriptions must cover the same scope.

One could identify an example of another definition proposed by the EU within the Proposal for a Directive of the European Parliament and of the Council on combating violence against women and domestic violence that refers to the production and dissemination of non-consensual deep porn[3] [54]. According to some estimations, deep fakes of a pornographic nature might constitute more than 90% of all deep fakes circulating on the internet, which clearly shows the scale of the problem [53] and explains the rationale behind including deep fakes into this particular directive in Recital 19:

> *The offence should also include the non-consensual production or manipulation, for instance by image editing, of material that makes it appear as though another person is engaged in sexual activities, insofar as the material is subsequently made accessible to a multitude of end-users, through information and communication technologies, without the consent of that person. Such production or manipulation should include the fabrication of 'deepfakes', where the material appreciably resembles an existing person, objects, places or other entities or events, depicting sexual activities of another person, and would falsely appear to others to be authentic or truthful.*

Regardless of the fact that the creation of deep porn materials should become a criminal offence and the limitation be of a contextual nature, inconsistency in the use of terms draws attention. Spelling discrepancies ('deepfakes' and 'deep fakes'[4]) are not as significant as the varying scope of definitions. It might be surprising that, only in this case, when the depiction of existing persons seems to be of importance due to the nature of deep porn [35], other elements being a part of the subjective aspect are also explicitly mentioned.

Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (DSA) [56] does not offer any legal definition of deep fakes but it apparently refers to that phenomenon in Article 35(1). While discussing 'mitigation of risks', DSA points to

3 —— Altered material of a sexual or pornographic nature, depicting people whose faces were superimposed on visual or audiovisual content [53].

4 —— With regard to spelling, one can also note the notation used by experts from the Panel for the Future of Science and Technology (STOA), who consistently used the term 'deep-fakes' in the report on the draft AI Act [55].

the obligations of providers of very large online platforms and very large online search engines, who should put in place reasonable, proportionate and effective mitigation measures. Such measures may include, where applicable:

> *k) ensuring that an item of information, whether it constitutes a generated or manipulated image, audio or video that appreciably resembles existing persons, objects, places or other entities or events and falsely appears to a person to be authentic or truthful, is distinguishable through prominent markings when presented on their online interfaces, and, in addition, providing an easy to use functionality which enables recipients of the service to indicate such information.*

In that case, attention should be drawn primarily to the extended subjective aspect, as the definition covers persons, objects, places or other entities or events. The issue of complementarity and potential strengthening of DSA provisions by the AI Act will be discussed later in the study.

A potential legal act that might in future also include a reference to deep fakes, due to their possible malicious applications in shaping political reality, influencing elections and causing risk of reputational harm to individuals [6], is the Proposal for a Regulation of the European Parliament and of the Council on the transparency and targeting of political advertising [57]. At the moment, this regulation does not directly refer to deep fakes. The table below summarises different definitions or references to deep fakes in EU legal acts.

Table 1. Different definitions and descriptions of deep fakes in the EU legal acts.

The AI Act proposal – European Commission, Article 52(3) (April 2021)	*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')*
The AI Act Proposal – European Parliament, Article 3(1) point 44d (June 2023)	*'deep fake' means manipulated or synthetic audio, image or video content that would falsely appear to be authentic or truthful, and which features depictions of persons appearing to say or do things they did not say or do, produced using AI techniques, including machine learning and deep learning*

The AI Act Proposal – European Parliament, Article 52(3) (June 2023)	*an AI system that generates or manipulates text, audio or visual content that would falsely appear to be authentic or truthful and which features depictions of people appearing to say or do things they did not say or do, without their consent ('deep fake')*
Directive on combating violence against women and domestic violence, Recital 19 (March 2022)	*'deepfakes', where the material appreciably resembles an existing person, objects, places or other entities or events, depicting sexual activities of another person, and would falsely appear to others to be authentic or truthful*
Regulation on a Single Market For Digital Services (DSA), Article 35(1k) (October 2022)	*an item of information, whether it constitutes a generated or manipulated image, audio or video that appreciably resembles existing persons, objects, places or other entities or events and falsely appears to a person to be authentic or truthful*

3. Qualification – specific risk category?

Any classification of deep fakes within a risk category should first take into account the possible uses of the technology that definitely may vary [12]. Deep fakes should not only be considered as a dangerous form of audio and visual manipulation, as there are many positive applications of the technology. They should not be described as something 'inherently morally wrong' and the technology itself should rather be considered 'neutral' [13, 27]. It is the use of deep fakes that gives them a certain dimension, and the objectives behind their creation or dissemination that put them into a specific context. The aforementioned elements of the definition do not comprise the aspect of contextuality, which often determines their harmfulness, and thus can be a key factor in risk assessment.

Strong emphasis on the negative uses of deep fakes has given them a bad reputation. There is a possibility that their excessive demonisation will lead to inappropriate risk assessment, or will undermine significant scientific and technological progress achieved with the use of deep fakes. Excessive interventionism, even if motivated by the protection of higher goods, can significantly limit technological development, and thus the competitiveness of the EU. Therefore any regulatory framework must be well-balanced. Discussing the positive uses of deep fakes is beyond the scope of this study, but it is worth noting that the term itself 'now carries negative connotations, potentially causing hesitancy or scepticism when discussing its legitimate research applications' [58].

Some authors believe that deep fakes are not a neutral technology, and that their history basically began with the creation of pornographic content, which clearly shows its original, highly disturbing objectives [59]. EUROPOL experts [60] estimate that most deep fakes are disseminated with malicious intent. Additionally, an intrinsic feature of deep fakes is that they increase confusion by blurring the boundaries between the authentic and the inauthentic and make it difficult to distinguish what is fact and what is fiction [59], significantly enhancing the potential for digital disinformation [25, 33]. Therefore, while assessing the potential applications of deep fakes, one should take into account the general negative consequences they cause within the information space, including undermining trust in information or the media [33, 61, 62].

The AI Act introduces a gradation of three basic risk categories: (i) unacceptable risk, (ii) high risk, and (iii) low or minimal risk. A detailed discussion on the legitimacy of such a division goes beyond the scope of this study, but the general idea of risk regulations is to prevent risk by reducing the probability of its occurrence [63]. It should be noted that in some respects the categorisation proposed in the AI Act is 'illusory and arbitrary' and does not apply to the entire 'AI lifecycle', which excludes or does not entirely cover the harmful forms of use of some systems [64]. The very general references (amended Recital 4 of the AI Act [24]) to the societal harm that some systems pose do not help to achieve clarity and certainty of the categorisation [65]. The AI Act does not provide a clear rationale for classifying deep fakes into any of the categories.

Pursuant to the AI Act, deep fakes were not qualified within the first two categories, so they should be automatically considered a low or minimal risk AI system. However, Title IV of the AI Act takes into account the specific risks of manipulation that some AI systems pose and thus introduces additional transparency obligations for specific AI systems. Deep fakes were enumerated among them and covered within the scope of the aforementioned Article 52(3) of the AI Act.

Deep fakes (and chatbots, pursuant to Article 52) must be treated as exceptions within the three-risk-categories system introduced by the AI Act [66] and might be classified as a 'specific risk' or 'limited risk' AI system [19, 20]. Therefore they form a separate quasi-category [67, 68].

Initially, in Recital 38 of the AI Act [4], which enumerates some high-risk systems, the Commission pointed out: *in view of the nature of the activities in question and the risks relating thereto, those high-risk AI*

systems should include in particular AI systems intended to be used by law enforcement authorities (...) to detect 'deep fakes'. According to the qualification made by the Commission, AI systems intended for use by law enforcement authorities to detect deep fakes were included in the list of High-Risk AI Systems (Annex III).

It should then be noted that originally deep fakes were classified into the third category or quasi-category (low or minimal risk, or specific or limited risk due to transparency obligations), while in the Commission's proposal deep fake detection systems were placed in the second category (high-risk). From the beginning, this qualification discrepancy gave rise to astonishment [21, 22, 69]. The misclassification was unconvincingly justified by the assumption that the former are used by the private sector, while the latter would be mainly in the hands of the public sector. Researchers from the European Parliamentary Research Service concluded: 'It is surveillance by the state individuals need protection from' [55]. This justification definitely loses to a practical approach to threats, primarily due to the malicious uses of deep fakes and the necessity to introduce efficient state countermeasures.

The European Parliament [24] effected a key change in this respect by making two deletions regarding deep fake detection systems – in Recital 38 and from the list of High-Risk AI Systems (Annex III) added to the AI Act. This is a direct and rational response to expert reservations and a common-sense approach – assigning a higher risk category to a technology that is supposed to protect against abuses with the use of technology classified in a lower risk category does not make sense. However, reservations may be made to paragraph 1 point 6d of Annex III, where the following are listed among high-risk AI systems: *AI systems intended to be used by or on behalf of law enforcement authorities, or by Union agencies, offices or bodies in support of law enforcement authorities to evaluate the reliability of evidence in the course of investigation or prosecution of criminal offences.* Deep fake detection systems could be indirectly included in the list [70], whereas their use in verifying the veracity of evidence has the potential for growth and is even recommended to ensure evidence integrity [60]. Experts consistently emphasise the importance of deep fake detection tools for counteracting deep fakes of a malicious nature, strengthening the capacity of law enforcement authorities, or protecting judicial proceedings [12, 60, 71]. These countermeasures will probably play an increasingly important role in the face of a growing number of crimes involving the malicious use of deep fakes (extortion, impersonation, financial fraud, forging evidence).[5] Separate issues are the effectiveness and credibility of the detection tools, as well as ensuring fair access to their

5 —— Deep fakes or their appearance in the information space have already been used during court proceedings to provide evidence (case in the UK during a custody battle), or to create a specific line of defense (the so-called 'deep fake defense') by claiming that evidence was fabricated [72], [73]. They also strengthen the 'liar's dividend', allowing depicted persons to claim that real content is in fact fake [6].

use, which should be an element of risk assessment that takes into account trustworthiness.

It is still necessary to return to the basic qualification of deep fakes in risk categories. Some researchers questioned the classification of deep fakes within the low or minimal risk category from the beginning, postulating their inclusion in the high-risk category [22] or reconsidering initial qualification [74]. The proposals were based on the correct assumption that counteracting deep fakes only through transparency obligations misses an important aspect of audiovisual content's manipulative capabilities. Mesarčík et al. [22] advocated for consistency in qualification rules for high-risk systems and accused the Commission of lacking a rationale in the case of deep fakes, the harmfulness of which might directly violate selected fundamental rights. In addition, they indicated the lack of a definition of inappropriate uses of deep fakes. All these objections seem valid, but difficult to grasp due to the high contextualisation of deep fakes and diversification of their applications.

A group of scientists conducting research on the harmful uses of deep fakes [13] indicated that 'manipulations may exhibit different risk levels and the risk level highly depends on the type of specific applications and somewhat subjectively depending on the actual use case'. This is an extremely important observation that relates in the first place to the various purposes behind the creation and dissemination of deep fakes and the contextuality of deep fakes as information carriers [75]. R. T. Toparlak [16] rightly noted that 'the wide range of applications means some deep fakes are going to be high-risk, while others are completely harmless'.

Theoretically, it is the objectives and the appropriate context of a particular deep fake that should determine its qualification into a risk category. In fact, they could be divided into many subgroups, depending on their form and purpose of use. This would of course give rise to problems of interpretation [76], which would have to be resolved on a case-by-case basis. However, the division would better reflect the specificity of the use of deep fakes and their destructive impact on the information ecosystem, individuals and society.

Assessing the harmfulness of deep fakes or the purposes of their creation and dissemination on a case-by-case basis seems rather unrealistic, which in turn undermines the legitimacy of the exception-based or multi-qualification risk system. Some authors [59] rightly noted that the scale of production of audio and visual materials is so large that it exceeds the verification capabilities of any institution, and

the verification itself would most likely have to be based on human review [21].

The answer to the question of whether deep fakes should qualify as a high-risk AI system is not clear. In light of descriptions of the high-risk AI systems category presented in the AI Act, one can have reasonable doubts whether deep fakes fail to meet at least some of the criteria. In the Explanatory Memorandum of the AI Act, it was noted that high-risk AI systems pose significant risks to the health and safety or fundamental rights of persons, which some deep fakes definitely do, including causing psychological harm to groups and individuals [77]. Deep fakes can also benefit from subliminal techniques [78] that are generally prohibited pursuant to Article 5(1a) of the AI Act. Problems may arise, however, in qualifying at which point a deep fake becomes a subliminal deep fake. Such difficulties may occur in the case of microtargeted video deep fakes based on facial resemblance and mimicry, which increase trustworthiness or self-enhancement among recipients [79–81].

The pillar on which the system the AI Act is built on is trustworthiness. Unfortunately, this system has some gaps as it mainly concentrates on the intended uses of specific AI systems and 'applies mandatory requirements for pre-defined domains of use', leaving some misuses and abuses unregulated [82]. Leaving deep fakes outside the scope of the high-risk category matches the general concept behind the AI Act risk assessment, but it does not take into account the fundamental malicious misuses of technology.

In the author's opinion, the reasonable solution for now would be to leave deep fakes within the low or minimal risk category with specific transparency obligations and distinguishing very concrete subgroups/exceptions for reclassification into the high-risk category or even imposing direct bans, which is needed in the case of deep porn [16]. Another issue is the fundamental effectiveness of the permissions, bans and transparency obligations, which will be discussed later.

From the point of view of strengthening social awareness and resilience, it is important to indicate why deep fakes give rise to threats – the AI Act might be the most appropriate place for the proper remarks. The European Parliament's Committee on the Internal Market and Consumer Protection and the European Parliament's Committee on Civil Liberties, Justice and Home Affairs [83] were clearly not sure about the qualification of deep fakes as a specific risk AI system. In a draft report from April 2022, it was proposed to add Recital 40a to the AI Act. It was supposed to clearly state:

> *Certain AI systems should at the same time be subject to transparency requirements and be classified as high-risk AI systems, given their potential to deceive and cause both individual and societal harm. In particular, AI systems that generate deep fakes representing existing persons have the potential to both manipulate the natural persons that are exposed to those deep fakes and harm the persons they are representing or misrepresenting, while AI systems that, based on limited human input, generate complex text such as news articles, opinion articles, novels, scripts, and scientific articles ('AI authors') have the potential to manipulate, deceive, or to expose natural persons to built-in biases or inaccuracies.*

This is a particularly interesting approach, questioning the Commission's initial qualification. Attention was rightly paid to the potential of deep fakes to deceive or cause harm to individuals and society. Such a comment – regardless of the final qualification of deep fakes – should appear within the AI Act to highlight the problem of manipulation, as well as the huge, often irreparable damages inflicted on individuals [84].

Interestingly, the German Bundesrat [85] was one of the few European chambers of parliaments to refer to the Commission's proposal in a resolution from September 2021 and touch upon the issue of deep fakes directly. It was rightly emphasised that deep fakes can manipulate public discourse in a covert manner, thereby exerting a significant influence on the process of individual and public opinion formation, and that they should not be treated as a side effect of the use of AI. It was suggested to consider deep fakes as a high-risk AI system and foreseen that this part of media law would have to be addressed properly by Member States since the AI Act does not cover that dimension properly [85].

It cannot be ruled out that in future deep fakes will become the subject of further thorough analyses and will be included in the list of high-risk AI systems. This type of evaluation will have to take into account, above all, the development of technology and its actual applications, for which permanent case study monitoring is necessary. First of all, it will be necessary to evaluate the validity and effectiveness of the introduced countermeasures. There is a high probability that transparency obligations alone will be insufficient to stop the vast majority of deep fakes of a malicious nature and that even moving to a higher risk category and becoming subject to strict obligations will not significantly change these negative trends. The remark from the Bundesrat in regard to the engagement of Member

States may actually indicate what path to combat deep fakes will become a future priority.

4. Transparency obligations and disclosure rules

The European Parliament's Committee on Legal Affairs [86], already at the beginning of 2021, indicated that deep fakes should be generally covered by disclosure rules, as they *could be used to blackmail, generate fake news reports, or erode public trust and influence public discourse; (...) such practices have the potential to destabilise countries, spreading disinformation and influencing elections.* The AI Act followed up on that assumption, though the regulation itself does not directly refer to the above-mentioned misuses of deep fakes.

Initially [4], it was proposed by the Commission that deep fakes would be classified as systems for which 'minimum transparency rules' would be required. This approach aroused justified controversy due to the threats associated with the presence of deep fakes in the information space. Mesarčík et al. [22] rightly pointed out that the proposed obligations lacked robustness and did not have the potential to significantly 'reduce the information asymmetry and thus allow the users (citizens) to combat the effects of deepfakes and still form informed and accurate opinions'.

The key to regulating the transparency obligations for deep fakes is Article 52(3) of the AI Act, which was fundamentally extended by the European Parliament. Initially [4], it contained an extremely general provision: *shall disclose that the content has been artificially generated or manipulated.* The amended version of Article 52(3) [24] introduces much more specific regulations that allow us to look at the solutions with cautious optimism:

> *shall disclose in an appropriate, timely, clear and visible manner that the content has been artificially generated or manipulated, as well as, whenever possible, the name of the natural or legal person that generated or manipulated it. Disclosure shall mean labelling the content in a way that informs that the content is inauthentic and that is clearly visible for the recipient of that content. To label the content, users shall take into account the generally acknowledged state of the art and relevant harmonised standards and specifications.*

Additionally, the European Parliament [24] has rightly added Article 52(3b), addressing some features of disclosure, and introduced special protection for vulnerable persons:

> *The information referred to in paragraphs 1 to 3 shall be provided to the natural persons at the latest at the time of the first interaction or exposure. It shall be accessible to vulnerable persons, such as persons with disabilities or children, complete, where relevant and appropriate, with intervention or flagging procedures for the exposed natural person taking into account the generally acknowledged state of the art and relevant harmonised standards and common specifications.*

This is definitely a step in the right direction and another of the significant and positive changes to the draft of the AI Act proposed by the European Parliament. The phrases 'appropriate, timely, clear and visible' seem to be of extreme importance, but it should be remembered that only standardisation processes allowing for the introduction of clear disclosure rules will enable final assessment of the adopted solutions and measuring their effectiveness in regard to some deep fakes (those that will be subject to any transparency obligations at all).

The Commission did not specify who would be the addressee of the disclosure [34]. The Parliament's amendments are more precise in this regard, even if they refer to the broad term of 'recipients'. In regard to deep fakes, transparency obligations are primarily meant to sensitise recipients and raise their awareness, or even serve to protect 'some right to reality grounded in fundamental rights' [23]. They are intended to show that recipients are dealing with fake content that does not represent reality – either distorting it in its entirety or falsifying it in order to mislead the audience [21]. The Explanatory Memorandum of the AI Act indicates that the obligation to disclose should allow recipients to make informed choices or step back from a given situation. The early-warning system is aimed at protecting recipients, their awareness and, to a large extent, trust in the information system. The positive impact of disclosure rules should then be considered mainly in the context of disinformation or media consistency. The rationale behind the provisions seems to be clear – deep fakes must be properly labelled due to their deceptive potential.

However, it should be considered whether transparency obligations will actually effectively protect recipients against disinformation. Expecting state or non-state actors with malicious goals to comply

with AI disclosure rules is obviously irrational. Rather, it should be assumed that transparency obligations will play a role in reducing the number of deep fakes circulating in the information space, especially those created by users equipped with unsophisticated software, but will not be a barrier for specialised actors.

In 2023 alone, deep fakes were successfully used in the US, Turkey and Germany, where they played a role in either influencing the election results or in fuelling current divisive issues. In Turkey, one of the opposition candidates in the presidential elections, Muharrem İnce, fell victim to deep porn and had to withdraw his candidature. İnce accused Russia of meddling in the Turkish elections [87]. In 2022, Russia used a deep fake video depicting the President of Ukraine, Volodymyr Zelensky, who was supposedly calling on his troops to surrender [88]. In the US, supporters of the Republican Party's rival candidates – Donald Trump and Ron DeSantis – continuously publish deep fake images and videos ridiculing their opponents [89]. US President Joe Biden is regularly the target of falsified information intended to damage his reputation, especially in the context of the 2024 US elections [90]. In Germany, deep fake videos depicting Canadian psychologist Jordan Peterson were disseminated to discredit the Minister of Foreign Affairs, Annalena Baerbock [91], while Minister of Economy Robert Habeck allegedly announced the closure of all outdoor swimming pools in response to incidents of violence [92]. The latter incident was intended to cause additional social unrest.

In 2023, a disturbing trend of using images of public figures to publish hate speech, anti-Semitic, racist or misogynistic content was observed. The voice of the popular actress Emma Watson was used to generate an audio deep fake in which she read fragments of Adolf Hitler's 'Mein Kampf' [93]. Journalists Joe Rogan and Ben Shapiro allegedly made homophobic and transphobic remarks [94].

Although most manipulations seem to be internally driven, the influence of external actors, including foreign countries, in cases of a strictly political nature, cannot be ruled out. The outreach and impact of content is generally multiplied by public willingness to share it, which mirrors the patterns of spreading disinformation due to injecting 'false but compelling information into a ready and willing information-sharing environment' by ordinary users [6].

In fact, transparency obligations in the form introduced by the AI Act could not be enforced in most of the cases mentioned above (assuming the applicability of the law due to jurisdiction). This results

directly from the intentions of its authors, which include, first and foremost, intentional and conscious misleading of recipients. As some deep fakes are created for the purposes of foreign information manipulation and interference, it should be assumed that the state and non-state actors involved in this practice will, for obvious reasons, not comply with any transparency obligations. In this context, simple technical solutions based on disclosure will be toothless [75, 95]. Therefore, the solutions proposed in the AI Act do not fully take into account the specificity of creating and disseminating deep fakes, the context of international politics and already known patterns of disinformation.

As a result, the transparency obligations 'will be applicable to only a small portion of deep fakes' [75]. The analysis by M. Veale and F. Z. Borgesius [23], who are quite critical of the way deep fakes were regulated in the AI Act, rightly pointed out that 'disclosure may only partially assist the subject', which in view of potential limitations on the effects of disclosure rules, may not be sufficient.

Protected goods must also include the personal rights of third parties whose image is the subject of the synthesis. Unfortunately, disclosure alone would not protect the subject/object of the depiction entirely. The organisation Access Now rightly pointed out that in many cases the 'transparency obligation will be insufficient to mitigate all risks associated with such applications' [96]. It needs to be clearly stated that in regard to deep porn, transparency obligations would not prevent the victimisation of depicted persons [75]. It is similar in cases of defamation or discrediting of individuals, when deep fakes can act as a catalyst for long-term negative emotions and associations. Research on the long-term consequences of exposure to fake news has shown that prior exposure increases the perceived accuracy of fake news [97]. Disclosure would not be able to stop these processes entirely. It would also not counteract the negative phenomenon of increased uncertainty in the case of exposure to fake content that might in turn undermine trust in the media, as proven by the experiment conducted by C. Vaccari and A. Chadwick [33]. The connection between disclosure and the actual reactions of recipients to AI-generated content could become the subject of research involving an evaluation of neural pathways and the possible outcomes of interference between two different messages – false information and disclosure of the falsehood.

It should also be noted that the AI Act imposes transparency obligations on 'users', while in the case of chatbots, it refers to 'providers'. Also in this regard, one may have concerns as to whether the transfer

of the burden to users is justified [20, 23], or even if provisions might be creating a 'legal loophole' [98]. A similar point was made by N. Helberger and N. Diakopoulos [68], who indicated that responsibility for the use of AI systems should lie primarily with providers, not users. The opposite would shift responsibility to end-users and disregard the potential risks of misusing certain systems.

Expanding the scope of provisions in regard to deep fakes might extend to potential additional legal obligations for providers [16]. EU regulations might oblige software providers to comply with fundamental rights and require further transparency [99], which in turn would add certainty to introduced solutions [98]. Moreover, the EU should understand the weaknesses of the AI Act in relation to counteracting deep fakes in order to consistently increase the legal regime in other areas. The postulated synergy effect between the AI Act and the DSA or Strengthened Code of Practice on Disinformation [21] seems to be a rational approach that takes into account various aspects of the negative impact of deep fakes. It should be emphasised again that precision, internal consistency and solution complementarity are necessary in this respect.

Another problem seems to be label parameters. Undoubtedly, markings should appear at the beginning of the material (pursuant to Article 52 point 3b of the AI Act [24], the information *shall be provided to the natural persons at the latest at the time of the first interaction or exposure*), though technical solutions might vary depending on the form of media used. In the case of video deep fakes, it seems advisable to disclose the fake nature of the content throughout playback in text form, so recipients are constantly aware that their interaction is based on AI activity. In the case of image deep fakes, it should be clearly and visibly stated in text form and disclosure should be an integral part of the image. In the case of audio deep fakes, it seems advisable to adjust disclosure to the length of the audio and the information should be read at least at the beginning and at the end of the display. Standardisation processes should take into account existing regulations and experiments on forms of disclosure. One interesting example is Bill S.B. 5152, adopted by the Washington State Senate in 2023[6] [100].

One might have reservations about the form of disclosure if only fragments of audio or visual content have been manipulated. It seems reasonable to ask whether, as a rule, the general pattern of AI disclosure should apply, or whether it should be modifiable and indicate which part of the material bears traces of AI interference [34]. In the author's opinion, it seems reasonable to label entire

6 —— Washington State Senate Bill on Defining synthetic media in campaigns for elective office, and providing relief for candidates and campaigns (S.B. 5152) [100] states: (4) It is an affirmative defense for any action brought under this section that the electioneering communication containing a synthetic media includes a disclosure stating, 'This (image/video/audio) has been manipulated,' in the following manner: a) For visual media, the text of the disclosure must appear in a size easily readable by the average viewer and no smaller than the largest font size of other text appearing in the visual media. If the visual media does not include any other text, the disclosure must appear in a size that is easily readable by the average viewer. For visual media that is a video, the disclosure must appear for the duration of the video; or (b) If the media consists of audio only, the disclosure must be read in a clearly spoken manner and in a pitch that can be easily heard by the average listener, at the beginning of the audio, at the end of the audio, and, if the audio is greater than two minutes in length, interspersed within the audio at intervals of not more than two minutes each.

content without any distinctions to avoid further manipulation and misleading of recipients.

The importance of appropriate labelling is emphasised by persistent low social awareness. Research conducted by Bitkom in Germany [102] shows that only 15% of respondents are able to explain what a deep fake is, and a mere 23% have basic knowledge on the subject. As many as 84% of respondents are in favour of marking deep fakes. Taking into account the very low number of respondents who are familiar with deep fakes, labelling must be adjusted to different kinds of audiences, which might be partly achieved by using simplified, concrete language.

A standardisation effort will be necessary in this regard. The Commission 'has begun to adopt a standardisation request which will provide a formal mandate to European standardisation organisations to develop standards under the AIA [AI Act]' [103]. Similarly, Article 82b(1) of the AI Act added by the European Parliament [24] indicates that the Commission shall develop, in consultation with the AI office, guidelines on the practical implementation of this Regulation, and in particular on the practical implementation of transparency obligations laid down in Article 52.

This area of research seems to be understudied and researchers need to enhance the outcome of standardisation processes. However, a very recent research study conducted by Dutch scientists dealing with deep fakes [104] is noteworthy, as it simulated the marking of video materials using three colours: green (veracity confirmed), yellow (veracity not confirmed), red (content containing false messages). The research results show that even such basic disclosure significantly increases scepticism among recipients and affects credibility assessments of the material. The researchers also tested the display time of the messages. Undoubtedly, such experiments must be repeated and modified in future to work out the best possible formula to measure when exactly labels should be displayed.

One could plausibly argue that even disclosure would not solve the problem of vulnerability to manipulation, or that the correlation between mere disclosure of using an AI system and increased protection of fundamental rights is relatively weak [55], but disclosure alone is a first step to protection and reduction of the negative effects of (some) deep fakes. An additional solution might be watermarking deep fake content [105], authenticating real content, or strengthening cyberliteracy to raise awareness among recipients.

5. Exceptions

Transparency obligations for deep fakes provide certain exceptions to the basic principles. The European Parliament [24] has made some significant changes to the Commission's proposal, also by extending and specifying the scope of exceptions. After amendment, Article 52(3a) states:

> *Paragraph 3* [transparency obligations] *shall not apply where the use of an AI system that generates or manipulates text, audio or visual content is authorised by law or if 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. Where the content forms part of an evidently creative, satirical, artistic or fictional cinematographic, video games visuals and analogous work or programme, transparency obligations set out in paragraph 3 are limited to disclosing of the existence of such generated or manipulated content in an appropriate clear and visible manner that does not hamper the display of the work and disclosing the applicable copyrights, where relevant. It shall also not prevent law enforcement authorities from using AI systems intended to detect deep fakes and prevent, investigate and prosecute criminal offences linked with their use.*

The exceptions therefore include two basic groups:

- authorisation by law (and in a later part – detection of deep fakes);
- exercise of the right to freedom of expression and the right to freedom of the arts and sciences that includes evidently creative, satirical, artistic or fictional cinematographic, video games visuals and analogous work or programme.

The former point does not seem to be controversial. It is the latter that has the potential to cause interpretation problems. Hertie School of Governance experts [75] predict that exceptions will open the door to creative manipulation and are 'likely to bring inconsistencies in practice'. The proposal rightly seeks to ensure high-level protection of the fundamental rights guaranteed by the EU Charter of Fundamental Rights, including freedom of expression (Article 11), or freedom of art and science (Article 13), but the practice may prove treacherous as the system of exceptions would possibly pave the way for exploitation.

Omission of the term 'timely' in Article 52(3a) of the AI Act, in comparison to Article 52(3) [24], leads to unnecessary problems of interpretation, especially since the legislator's intention was apparently to approximate the provisions on non-exceptions and exceptions within this particular AI system. Based on very general formulations, it is difficult to determine what disclosure would actually look like in the case of exceptions. The distinction itself opens up room for manipulation and misinterpretation.

In an increasing number of cases, legislators have prohibited or limited the use of deep fakes, but they have also allowed significant exceptions in the form of obvious or evident satire or parody (of a 'demonstrably' fake nature) [25, 106–108]. This 'obvious' or 'evident' nature may be debatable and would have to be assessed on a case-by-case basis because it might depend on contextualisation as well as the cognitive abilities, media knowledge, or social and political awareness of recipients.

Unfortunately, overusing the legal exemptions could be seen as a useful tool to circumvent the restrictions. Deep fakes are described as a phenomenon that might benefit from the 'just joking' excuse, making it possible to smuggle illegal content or manipulate the audience 'under the guise of humour', which might even lead to the 'weaponisation of humour' [109]. Satirical context has already been shown to function as 'a cover for spreading' extremist ideologies [25] with respect to fake news. At the same time, the fight against deep fakes might also be used to justify suppressing freedom of speech. This is especially important in the case of non-democratic countries that hide their censorship tendencies under the guise of protecting social stability [15].

While deep fakes can be successfully used to create content that is critical of the authorities, the limits of satire are hard to grasp, especially since the boundaries between satire and harmful content are increasingly blurred. Difficulties also arise when 'satire is transferred out of its original context', is 'no longer recognisable' due to high synthesis quality, or is not recognised by recipients [110].

One potential solution could be to treat all deep fakes in the same way with respect to transparency obligations. If the satirical or parodic nature of the material is obvious, disclosing the use of AI and appropriately flagging the fake content should not be a problem and standardised transparency obligations would help to protect recipients. It seems reasonable to refer to fundamental rights, including freedom of speech, while noting that the requirements regarding

transparency obligations will not violate these rights. Labelling AI-manipulated audio or visual content should be seen as a standard rather than an arduous obligation.

6. Conclusions

Ongoing work on the AI Act in regard to deep fakes gives hope for more robust protection of EU citizens against AI manipulation. It should be emphasised that the amendments introduced by the European Parliament would slightly increase the ability of the EU to counteract the negative effects of deep fakes. The imperfection of the solutions results to a large extent from general legislative difficulties related to the creation and dissemination of deep fakes, the specificity of deep fakes and the complexity of the challenges they create for democratic systems, societies and individuals, but it is also due to an internal lack of EU coherence or precision.

The European Parliament has already introduced numerous positive changes to the Commission's proposal for the AI Act, thus addressing some of the critical analyses by experts. However, this does not mean that the regulation is free of deficits in its current form. A detailed analysis of the Commission's proposal and European Parliament amendments in regard to deep fakes allows us to draw a number of conclusions and identify reservations that should be considered in future revisions of the AI Act.

1. The level of expectations should be adjusted to the AI Act's true capabilities to influence reality, especially since the provisions on deep fakes are not the key element of the regulation and the level of protection it offers against them is basic at best. The authors of a significant portion of deep fakes will neither comply with transparency obligations nor care about the risk categories [23, 75]. As emphasised, the vast majority of deep fakes comprises non-consensual pornography (deep porn). Such materials, due to their specific nature and manner of dissemination, will never be subject to any disclosure rules. In this context, it is necessary to implement stricter provisions aimed at protecting individuals, in particular women, against the deployment of gender-based violence, exploitation, humiliation, or harassment. The European Parliament resolution, containing recommendations to the Commission on combating gender-based violence [54], paves the way for further actions. This might be achieved through additional countermeasures, including putting pressure on platforms that enable the dissemination of such content,

which the DSA fortunately already does. In the case of most non-consensual malicious deep fakes (including deep porn), the basic idea of creating and disseminating deep fakes violates the law, even if the relevant provisions are derived from civil, tort, criminal or copyright law. Member States have to reconsider how to make these provisions more efficient. Therefore, the role that the AI Act would play in combating deep porn would be significantly reduced and one should not expect a breakthrough. The proposed transparency obligations seem to be appropriate to regulate a small portion of deep fakes appearing in the information space. This applies not only to deep porn but also to some disinformation activities that might be driven by foreign information manipulation and interference.

2. Hertie School of Governance experts [75] rightly pointed out that the AI Act offers the 'false promise of transparent deep fakes'. Disclosure rules give the illusory belief that revealing the false nature of content (if it gets done at all) will lead to the elimination of the negative effects of creating and disseminating deep fakes. It will not. The problem with deep porn or discrediting materials is the non-consensual use of someone else's image and the psychological and reputational harm it creates [26]. Even if disclosure rules are applied to non-consensual deep fakes (especially deep porn), the negative effects leading to psychological harm will not be eliminated. Many women have been victims of non-consensual pornography and have reported severe psychological effects, including discomfort while using social media, depression, anxiety or trauma [26, 96, 111]. Similar consequences can be measured with respect to false content of a discrediting nature since malicious deep fakes can cause reputational harm and thus have long-lasting repercussions on the psychological well-being or professional prospects of the depicted individuals [112].

3. The EU must ensure internal coherence, in particular in regard to the proposed definitions and descriptions of deep fakes. Therefore, there should be absolutely no internal discrepancies within the AI Act or between different legal acts proposed by the EU. The certainty of the law, its interpretation and enforcement must be an asset of EU legislative activity. The AI Act may set a common standard, to which subsequent legal acts will refer. Deep fakes must be unambiguously defined, and the definition must clearly include, among others, the scope of the form of deep fakes (typological aspect) and the subjects/objects to which deep fakes refer (subjective aspect). That applies mainly

to discrepancies between Recital 70, Article 3(1) point 44d and Article 52(3) of the AI Act. In the author's opinion, it is necessary to extend the scope of the definition to 'persons, objects, places or other entities or events', as well as to reconsider the potential omission of deep fakes in textual form.

4. Transparency obligations are not a universal solution. J. Habgood-Coote [95] may be right in pointing out that a significant number of researchers is guided by 'technochauvinism' or 'techno-fixation', assuming that the problem of deep fakes can be solved with the use of technological tools. It might be better to qualify deep fakes 'as a social problem about the management of our practices for producing and receiving recordings' [95]. At the same time, technological solutions can at least reduce negative trends, acting as a deterrent. That is why it is so important to find a balance between various ways of counteracting the harmful uses of deep fakes and their negative consequences, which might include disclosure, watermarking, content authentication, or strengthening cyberliteracy [12]. Even if it would help to eliminate only a small number of deep fakes, the AI Act should be seen as a step in the right direction, but it needs to be supplemented with further regulatory and non-regulatory efforts from the EU to strengthen social resilience, also by enhancing cyberliteracy. Again, omitting the critical reference to the specific risks that deep fakes pose overlooks a significant aspect of raising awareness through the AI Act.

5. If transparency and disclosure are to introduce a reasonable level of protection, it is necessary to tighten the system to prevent possible attempts to circumvent the obligations. It is advisable to reconsider sealing the system of exceptions to full disclosure rules. The assumption that the satirical or parody nature of the material is 'obvious' or 'evident' is based on a misconception about the high level of cognitive and analytical skills[7] among recipients of deep fakes [110]. Satire and parody can be successfully used to bypass some safeguards in order to smuggle sophisticated political manipulation and thus influence the audience. It will also be crucial to develop the practice for disclosure rules, which requires standardisation processes and empirical research to measure the effectiveness of different solutions. Ongoing work by Dobber et al. [104] as well as solutions introduced in the US might serve as an example. The EU should closely monitor regulatory efforts in other countries to either use the labelling patterns for standardisation processes, or even introduce concrete provisions within the AI Act. The transparency

7 — J. Langa [108], commenting on the provisions introduced in the US, refers to the notion of 'reasonable person' that 'realizes that a deepfake is satirical or parodical' and thus cannot be deceived. The term seems to be vague and the highly deceptive nature of deep fakes (especially video deep fakes) has been proven many times.

obligations might be complemented by imposing additional obligations on providers and manufacturers. Although this will not eliminate non-consensual deep fakes of a harmful nature, it will limit their effects and the amount of manipulated and unmarked content by making it more difficult for non-specialised users to create deep fakes [98].

6. The AI Act does not impose any special obligations on digital platforms in regard to the creation and dissemination of deep fakes. It can be argued that such solutions are found in other acts introduced by the EU, but the lack of an internal connection does not directly indicate the specific purpose and complementarity of measures counteracting deep fakes [75]. According to some experts [12], 'distribution and consumption patterns pose larger threats to democracy and society than the fake content itself'. It might be advisable to concentrate on prevention by delimiting the applications of technology, also through ethical norms [113], and reducing dissemination capabilities. The Centre for Data Innovations [114] suggested 'nimbler soft law approaches' to 'supplement adjustments to the AI Act and the Directive on Gender-based Violence' by working closely with industry and encouraging self-regulatory efforts to counteract non-consensual pornography. That would definitely fit with the idea of reducing the impact of deep fakes by restricting amplification of the content through online platforms [76], which applies not only to deep porn but also to other types of deep fakes, including those of an intrinsically political nature.

7. The European Parliament's amendments in regard to deep fake detection systems in the form of deletion from the list of high-risk AI systems should be assessed positively, primarily due to an initial erroneous discrepancy in the risk assessment between deep fakes and the technological measures that are intended to protect against them. The potential problem with the provision included in paragraph 1 point 6d in Annex III of the AI Act, which might lead to the indirect inclusion of deep fake detection systems in the high-risk AI systems list [70], might pose interpretational problems and should be clarified at a later stage in the negotiations.

8. It still seems controversial that deep fakes are not qualified to the category of high-risk AI systems, especially because the AI Act provides some rationale for reclassification. The potential solution might be to single out those deep fakes that pose

a greater threat to specific subjects for special protection and transfer them to a higher category, describing the scope of their harmfulness in a clear and precise manner to leave no room for misinterpretation (e.g. introducing additional protection for candidates before an election), or introducing a complete ban on their creation and dissemination (e.g. deep porn). The omission of the contextual aspect while assessing the risk posed by deep fakes can be assessed negatively. Reference to the harmful uses of deep fakes and detrimental effects they cause should be added in one of the recitals, which might also be extended by broader and more-detailed reference to the systemic and societal harms that AI systems might pose [65, 115]. At the same time, it should be remembered that even moving to a higher risk category will not be a universal solution or eliminate the basic problem related to the spread of some deep fakes of a malicious nature, since they are not subject to any rules.

9. The fundamental problem with the emergence of deep fakes in the information space is not a complete lack of regulation. In many cases, deep fakes of a malicious nature are directly or indirectly prohibited by law, and victims can pursue their rights in court. The problem, however, is enforcement of existing provisions [59]. The AI Act would not change this situation drastically, and some may rightly accuse the regulation of failing to impose sanctions for non-compliance with the transparency obligations [74]. These can easily be derived from other EU legal acts, including the DSA, which obliges platforms to inform users about the deceptive or manipulative nature of content [116]. Pursuant to the DSA, non-compliance can be sanctioned by up to 6% of annual worldwide turnover. The AI Act should complement these solutions, particularly with respect to authors of deep fakes and AI system providers [21], which currently is not the case. Although numerous researchers have pointed out that identifying perpetrators is problematic (the basic problem with attribution), the AI Act might add another source of pressure and mobilise law enforcement authorities and policymakers to deal with the problem [25, 98].

References

[1] J.Laux, S. Wachter, B. Mittelstadt, "Trustworthy artificial intelligence and the European Union AI act: On the conflation of trustworthiness and acceptability of risk," *SSRN Electronic Journal*, 2023, doi: 10.2139/ssrn.4230294.

[2] P. Maham, S. Küspert, *Governing General Purpose AI.* Stiftung Neue Verantwortung, Berlin, 2023.

[3] J. Schuett, "Risk Management in the Artificial Intelligence Act," *European Journal of Risk Regulation*, pp. 1–19, 2023, doi:10.1017/err.2023.1.

[4] European Commission, *Proposal for Regulation of the European Parliament and the Council laying down harmonized rules on Artificial Intelligence (Artificial Intelligence Act) and amending certain union legislative acts.* COM(2021) 206 final. 2021. [Online]. Available: https://eur-lex.europa.eu/resource.html?uri=cellar:e0649735-a372-11eb-9585-01aa75ed71a1.0001.02/DOC_1&format=PDF. [Accessed: Sep. 13, 2023].

[5] T. Brooks, G. Princess, J. Heatley, J. Jeremy, K. Scott, et al., *Increasing Threats of Deepfake Identities,* U.S. Department of Homeland Security, 2019. [Online]. Available: https://www.dhs.gov/sites/default/files/publications/increasing_threats_of_deepfake_identities_0.pdf. [Accessed: Sep. 13, 2023].

[6] B. Chesney, D. Citron, "Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security," *California Law Review*, vol. 107, no. 18, pp. 1753–1820, 2019, doi: 10.15779/Z38RV0D15J

[7] I. Dąbrowska, "Deepfake – nowy wymiar internetowej manipulacji," *Zarządzanie Mediami*, vol. 8, no. 2, pp. 89–101, 2020, doi: 10.4467/23540214ZM.20.024.11803.

[8] D. L. Byman, C. Gao, C. Meserole, *Deepfakes and international conflict*. Washington: The Brookings Institution, 2023.

[9] M. Pawelec, "Deepfakes and Democracy (Theory): How Synthetic Audio-Visual Media for Disinformation and Hate Speech Threaten Core Democratic Functions," *Digital Society*, vol. 2, no. 2, 2022, doi: 10.1007/s44206-022-00010-6.

[10] M. Mustak, J. Salminen, M. Mäntymäki, A. Rahman, Y. K. Dwivedi, "Deepfakes: Deceptions, mitigations, and opportunities," *Journal of Business Research*, vol. 154, 2023, doi: 10.1016/j.jbusres.2022.113368.

[11] E. Pashentsev, "The Malicious Use of Deepfakes Against Psychological Security and Political Stability," in *The Palgrave Handbook of Malicious Use of AI and Psychological Security*, E. Pashentsev, Ed. London: Palgrave Macmillan, Cham, 2023, pp. 47–80.

[12] H. Farid, H.-J. Schindler, *Deep Fakes. On the Threat of Deep Fakes to Democracy and Society*. Berlin: Konrad Adenauer Stiftung, 2020.

[13] F. Juefei-Xu, R. Wang, Y. Huang, Q. Guo, L. Ma, et al., "Countering Malicious DeepFakes: Survey, Battleground, and Horizon," *International Journal of Computer Vision*, vol. 130, no. 7, pp. 1678–1734, 2022, doi: 10.1007/s11263-022-01606-8.

[14] J. Ice, "Defamatory Political Deepfakes and the First Amendment," *Case Western Reserve Law Review*, vol. 70, no. 2, pp. 417–455, 2019.

[15] E. Hine, L. Floridi, "New deepfake regulations in China are a tool for social stability, but at what cost?," *Nature Machine Intelligence*, vol. 4, no. 7, pp. 608–610, 2022.,doi: 10.1038/s42256-022-00513-4.

[16] R. T. Toparlak, "Criminalising Pornographic Deep Fakes: A Gender-Specific Inspection of Image-Based Sexual Abuse," *SciencesPo Law School The 10th Graduate Conference*, 2022. [Online]. Available: https://www.sciencespo.fr/public/chaire-numerique/wp-content/uploads/2022/06/3a-Toparlak_Criminalising-Pornographic-Deep-Fakes.pdf. [Accessed: Sep. 13, 2023].

[17] E. Meskys, A. Liaudanskass, J. Kalpokiene, P. Jurcy, "Regulating Deep-Fakes: Legal and Ethical Considerations," *Journal of Intellectual Property Law & Practice*, vol. 15, no. 1, pp. 24–31, 2020, doi: 10.1093/jiplp/jpz167.

[18] K. Mamak, "Categories of Fake News from the Perspective of Social Harmfulness," in *Integrity of Scientific Research: Fraud, Misconduct and Fake News in the Academic, Medical and Social Environment*, J. Faintuch, S. Faintuch, Ed. Springer, pp. 351–357, 2022, doi: 10.1007/978-3-030-99680-2_35.

[19] M. Kop, "EU Artificial Intelligence Act: The European Approach to AI," *Transatlantic Antitrust and IPR Developments*, vol. 2, 2021. [Online]. Available: https://law.stanford.edu/wp-content/uploads/2021/09/2021-09-28-EU-Artificial-Intelligence-Act-The-European-Approach-to-AI.pdf. [Accessed: Sep. 13, 2023].

[20] L. Edwards, *The EU AI Act: a summary of its significance and scope*, Ada Lovelace Institute, 2022. [Online]. Available: https://www.adalovelaceinstitute.org/wp-content/uploads/2022/04/Expert-explainer-The-EU-AI-Act-11-April-2022.pdf. [Accessed: Sep. 13, 2023].

[21] A. Fernandez, "'Deep fakes': disentangling terms in the proposed EU Artificial Intelligence Act," *UFITA Archiv für Medienrecht und Medienwissenschaft*, vol. 85, no. 2, pp. 392–433, 2021, doi: 10.5771/2568-9185-2021-2-392.

[22] M. Mesarčík, S. Solarova, J. Podroužek, M. Bielikova, *Stance on The Proposal for a Regulation Laying Down Harmonised Rules on Artificial Intelligence – Artificial Intelligence Act*, Kempelen Institute of Intelligent Technologies, 2021, doi: 10.31235/osf.io/yzfg8.

[23] M. Veale, F. Z. Borgesius, "Demystifying the Draft EU Artificial Intelligence Act – Analysing the good, the bad, and the unclear elements of the proposed approach," *Computer Law Review International*, vol. 22, no. 4, pp. 97–112, 2021, doi: 10.9785/cri-2021-220402.

[24] European Parliament, *Amendments adopted by the European Parliament on 14 June 2023 on the proposal for a regulation of the European Parliament and of the Council on laying down harmonised rules on artificial intelligence (Artificial Intelligence Act) and amending certain Union legislative acts.* (COM(2021)0206 – C9-0146/2021 – 2021/0106(COD))1. P9_TA(2023)0236. 2023. [Online]. Available: https://www.europarl.europa.eu/doceo/document/TA-9-2023-0236_EN.pdf. [Accessed: Sep. 13, 2023].

[25] M. van Huijstee, P. van Boheemen, D. Das, L. Nierling, J. Jahnel et al., *Tackling deep-fakes in European policy*, European Parliamentary Research Service, Brussels, 2021.

[26] R. Delfino, "Pornographic Deepfakes: The Case for Federal Criminalization of Revenge Porn's Next Tragic Act," *Fordham Law Review*, vol. 88, no. 3, pp. 887–938, 2019, doi: 10.2139/ssrn.3341593.

[27] A. De Ruiter, "The Distinct Wrong of Deepfakes," *Philosophy & Technolology*, vol. 34, pp. 1311–1332, 2021, doi: 10.1007/s13347-021-00459-2.

[28] H. Farid, "Creating, Using, Misusing, and Detecting Deep Fakes," *Journal of Online Trust and Safety*, vol. 1, no. 4, pp. 1–33, 2022, doi: 10.54501/jots.v1i4.56.

[29] A. Satariano, P. Mozur, *The People Onscreen Are Fake. The Disinformation Is Real,* 2023. [Online]. Available: https://www.nytimes.com/2023/02/07/technology/artificial-intelligence-training-deepfake.html. [Accessed: Sep. 13, 2023].

[30] T. Weikmann, S. Lecheler, "Cutting through the Hype: Understanding the Implications of Deepfakes for the Fact-Checking Actor-Network," *Digital Journalism*, 2023, doi: 10.1080/21670811.2023.2194665.

[31] M. Westerlund, "The Emergence of Deepfake Technology: A Review," *Technology Innovation Management Review*, vol. 9, no. 11, pp. 39–52, 2019, doi: 10.22215/timreview/1282.

[32] S. J. Nightingale, H. Farid, "AI-synthesized faces are indistinguishable from real faces and more trustworthy," *Proceedings of the National Academy of Sciences,* vol. 119, no. 8, 2022, doi: 10.1073/pnas.2120481119.

[33] C. Vaccari, A. Chadwick, "Deepfakes and Disinformation: Exploring the Impact of Synthetic Political Video on Deception, Uncertainty, and Trust in News," *Social Media + Society*, vol. 6, no. 1, 2020, doi: 10.1177/2056305120903408.

[34] B. van der Sloot, Y. Wagensveld, "Deepfakes: regulatory challenges for the synthetic society," *Computer Law & Security Review*, vol. 46, 2022, doi: 10.1016/j.clsr.2022.105716.

[35] C. Okolie, "Artificial Intelligence-Altered Videos (Deepfakes), Image-Based Sexual Abuse, and Data Privacy Concerns," *Journal of International Women's Studies*, vol. 25, no. 2, 2023. [Online] Available: https://vc.bridgew.edu/jiws/vol25/iss2/11. [Accessed: Sep. 13, 2023].

[36] L. Hurst, *How a fake image of a Pentagon explosion shared on Twitter caused a real dip on Wall Street*, 2023. [Online]. Available: https://www.euronews.com/next/2023/05/23/fake-news-about-an-explosion-at-the-pentagon-spreads-on-verified-accounts-on-twitter. [Accessed: Sep. 13, 2023].

[37] C. Öhman, "The identification game: deepfakes and the epistemic limits of identity," Synthese, vol. 200, no. 4, 2022, doi: 10.1007/s11229-022-03798-5.

[38] A. Kleemann, "Deepfakes – Wenn wir unseren Augen und Ohren nicht mehr trauen können," *SWP-Aktuell*, vol. 43, 2023, doi: 10.18449/2023A43.

[39] K. Eckart, *A growing problem of 'deepfake geography': How AI falsifies satellite images*, 2021. [Online]. Available: https://www.washington.edu/news/2021/04/21/a-growing-problem-of-deepfake-geography-how-ai-falsifies-satellite-images. [Accessed: Sep. 13, 2023].

[40] B. Zhao, S. Zhang, C. Xu, Y. Sun, C. Deng, "Deep fake geography? When geospatial data encounter Artificial Intelligence," *Cartography and Geographic Information Science*, vol. 48, no. 4, pp. 338–352, 2021, doi: 10.1080/15230406.2021.1910075.

[41] K. Hiebert, Democracies Are Dangerously Unprepared for Deepfakes, 2022. [Online]. Available: https://www.cigionline.org/articles/democracies-are-dangerously-unprepared-for-deepfakes. [Accessed: Sep. 13, 2023].

[42] European Parliament's Committee on Industry, Research and Energy, *Opinion on 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,* 2022. [Online]. Available: https://www.europarl.europa.eu/doceo/document/ITRE-AD-719801_EN.pdf. [Accessed: Sep. 13, 2023].

[43] J. Bateman, *Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios*. Washington: Carnegie Endowment for International Peace, 2020.

[44] K. Giles, K. Hartmann, M. Mustaffa, *The Role of Deepfakes in Malign Influence Campaigns*. Riga: NATO Strategic Communications Centre of Excellence, 2019.

[45] J. R. Allen, D. M. West (2020). *The Brookings glossary of AI and emerging technologies*. [Online]. Available: https://www.brookings.edu/articles/the-brookings-glossary-of-ai-and-emerging-technologies. [Accessed: Sep. 13, 2023].

[46] T. Hwang, *Deepfakes: A Grounded Threat Assessment*. Washington: Center for Security and Emerging Technology, 2020.

[47] R. Mattioli, A. Malatras, *Identifying Emerging Cyber Security Threats and Challenges for 2030.* Athens: ENISA, 2023.

[48] M. Atleson, *Combatting Online Harms Through Innovation.* Federal Trade Commission Report to Congress, 2022. [Online] Available: https://www.ftc.gov/system/files/ftc_gov/pdf/Combatting%20Online%20Harms%20Through%20Innovation%3B%20Federal%20Trade%20Commission%20Report%20to%20Congress.pdf [Accessed: Sep. 13, 2023].

[49] A. J. Goldstein, G. Sastry, M. Musser, R. DiResta, M. Gentzel, et al., *Generative Language Models and Automated Influence Operations: Emerging Threats and Potential Mitigations.* 2023. [Online]. Available: arxiv.org/abs/2301.04246. [Accessed: Sep. 13, 2023].

[50] Z. Khanjani, G. Watson, V. P. Janeja, "Audio deepfakes: A survey," *Frontiers in Big Data*, vol. 5, 2023, doi: 10.3389/fdata.2022.1001063.

[51] J. Pu, Z. Sarwar, S. M. Abdullah, A. Rehman, Y. Kim et al., "Deepfake Text Detection: Limitations and Opportunities," *IEEE Symposium on Security and Privacy (SP),* 2023.,doi: 10.1109/sp46215.2023.10179387.

[52] European Parliament's Committee on Culture and Education, *Opinion on 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*. 2022. [Online]. Available: https://www.europarl.europa.eu/doceo/document/CULT-AD-719637_EN.pdf. [Accessed: Sep. 13, 2023].

[53] C. Rigotti, C. McGlynn, "Towards an EU criminal law on violence against women: The ambitions and limitations of the Commission's proposal to criminalise image-based sexual abuse," *New Journal of European Criminal Law*, vol. 13, no. 4, pp. 452–477, 2022, doi: 10.1177/20322844221140713.

[54] European Commission, *Directive on combating violence against women and domestic violence. Proposal for a Directive of the European Parliament and of the Council on combating violence against women and domestic violence'.* COM(2022) 105 final. 2022.

[55] I. Georgieva, T. Timan, M. Hoekstra, *Regulatory divergences in the draft AI act.* Brussels: Scientific Foresight Unit (STOA), 2022.

[56] European Union, *Regulation (EU) 2022/2065 of the European Parliament and of the Council of 19 October 2022 on a Single Market For Digital Services and amending Directive 2000/31/EC (Digital Services Act).* 2022.

[57] European Commission, *Proposal for a Regulation of the European Parliament and of the Council on the transparency and targeting of political advertising*. COM(2021) 731 final. 2021.

[58] C. Becker, R. Laycock, "Embracing deepfakes and AI-generated images in neuroscience research," *European Journal of Neuroscience*, vol. 58, no. 3, pp. 2657–2661, 2023, doi: 10.1111/ejn.16052.

[59] B. van der Sloot, Y. Wagensveld, B. J. Koops, *Deepfakes: The Legal Challenges of the Synthetic Society,* Tilburg Institute for Law, Technology, and Society, 2021, doi: 10.1016/j.clsr.2022.105716

[60] Europol, *Facing reality? Law enforcement and the challenge of deepfakes, an observatory report from the Europol Innovation Lab*. Brussels: Publications Office of the European Union, 2022.

[61] N. Schick, *Deep Fakes and the Infocalypse*. London: Octopus Books, 2020.

[62] J. Ternovski, J. Kalla, P. M. Aronow, "Deepfake Warnings for Political Videos Increase Disbelief but Do Not Improve Discernment: Evidence from Two Experiments," *OSF Preprints*, [Online]. Available: https://osf.io/dta97. [Accessed: Sep. 13, 2023].

[63] J. De Cooman, "Humpty Dumpty and High-Risk AI Systems: The Ratione Materiae Dimension of the Proposal for an EU Artificial Intelligence Act," *Market and Competition Law Review*, vol. 6, no. 1, 2022, doi: 10.34632/mclawreview.2022.11304.

[64] L. Edwards, "Regulating AI in Europe: four problems and four solutions", Ada Lovelace Institute, 2022. [Online] Available: https://www.adalovelaceinstitute.org/wp-content/uploads/2022/03/Expert-opinion-Lilian-Edwards-Regulating-AI-in-Europe.pdf. [Accessed: Sep. 13, 2023].

[65] Ada Lovelace Institute, *People, risk and the unique requirements of AI,* 2022. [Online]. Available: https://www.adalovelaceinstitute.org/wp-content/uploads/2022/03/Policy-briefing-People-risk-and-the-unique-requirements-of-AI-18-recommendations-to-strengthen-the-EU-AI-Act.pdf. [Accessed: Sep. 13, 2023].

[66] Commission Staff Working Document, *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.* 2021. [Online]. Available: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=celex%3A52021sc0084. [Accessed: Sep. 13, 2023].

[67] W. Wahlster, C. Winterhalter, *German Standardization Roadmap on Artificial Intelligence*, DIN, DKE, 2022. [Online]. Available: https://www.din.de/resource/blob/772610/e96c34dd6b12900ea75b460538805349/normungsroadmap-en-data.pdf. [Accessed: Sep. 13, 2023].

[68] N. Helberger, N. Diakopoulos, "ChatGPT and the AI Act," *Internet Policy Review*, vol. 12, no. 1, 2023, doi: 10.14763/2023.1.1682.

[69] T. Mahler, "Between risk management and proportionality: The risk-based approach in the EU's Artificial Intelligence Act Proposal," *Nordic Yearbook of Law and Informatics,* pp. 247–270, 2022, doi: org/10.53292/208f5901.38a67238.

[70] M. C. Sanchez, *Deep fakes: the media and the legal system is under threat*, 2023. [Online]. Available: https://www.lexology.com/library/detail.aspx?g=e4e835cb-f3d1-416e-81b8-81eefe426cf4. [Accessed: Sep. 13, 2023].

[71] T. T. Nguyen, Q. V. H. Nguyen, D. T. Nguyen, D. T. Nguyen, T. Huynh-The et al., "Deep Learning for Deepfakes Creation and Detection: A Survey," *SSRN Electronic Journal,* 2022, doi: 10.2139/ssrn.4030341.

[72] V. Cianciaglini, C. Gibson, D. Sancho, O. McCarthy, M. Eira et al., *Malicious Uses and Abuses of Artificial Intelligence*, United Nations Interregional Crime and Justice Research Institute, Europol's European Cybercrime Centre, 2020. [Online]. Available: https://www.europol.europa.eu/cms/sites/default/files/documents/malicious_uses_and_abuses_of_artificial_intelligence_europol.pdf. [Accessed: Sep. 13, 2023].

[73] R. Delfino, "The Deepfake Defense-Exploring the Limits of the Law and Ethical Norms in Protecting Legal Proceedings from Lying Lawyers," Preview Ohio State Law Journal, vol. 84, 2023, doi: 10.2139/ssrn.4355140.

[74] L. Holbrook (2023). *The EU Artificial Intelligence Act and its Human Rights Limitations*. [Online]. Available: https://ohrh.law.ox.ac.uk/the-eu-artificial-intelligence-act-and-its-human-rights-limitations. [Accessed: Sep. 13, 2023].

[75] Centre for Digital Governance, *The false promise of transparent deep fakes: How transparency obligations in the draft AI Act fail to deal with the threat of disinformation and image-based sexual abuse*, Hertie School, 2022. [Online]. Available: https://www.hertie-school.org/en/digital-governance/research/blog/detail/content/the-false-promise-of-transparent-deep-fakes-how-transparency-obligations-in-the-draft-ai-act-fail-to-deal-with-the-threat-of-disinformation-and-image-based-sexual-abuse. [Accessed: Sep. 13, 2023].

[76] K. Nagumotu, "Deep fakes are taking over social media: can the law keep up?," *Intellectual Property Law Review*, vol. 62, no. 2, pp. 102–146, 2022.

[77] M. Ebers, V. R. S. Hoch, F. Rosenkranz, H. Ruschemeier, B. Steinrötter, "The European Commission's Proposal for an Artificial Intelligence Act–A Critical Assessment by Members of the Robotics and AI Law Society (RAILS)," *Multidisciplinary Scientific Journal,* vol. 4, no. 4, pp. 589–603, 2021, doi: 10.3390/j4040043.

[78] R. J. Neuwirth, "Prohibited artificial intelligence practices in the proposed EU artificial intelligence act (AIA)," *Computer Law & Security Review*, vol. 48, 2023, doi: 10.1016/j.clsr.2023.105798.

[79] J. B. Finke, M. F. Larra, M. U. Merz, H. Schächinger, "Startling similarity: Effects of facial self-resemblance and familiarity on the processing of emotional faces," *PLOS ONE*, vol. 12, no. 12, 2017, doi 10.1371/journal.pone.0189028.

[80] T. Dobber, N. Metoui, D. Trilling, N. Helberger, C. de Vreese, "Do (Microtargeted) Deepfakes Have Real Effects on Political Attitudes?," *The International Journal of Press/Politics*, vol. 26, no. 1, pp. 69–91, 2021, doi: 10.1177/1940161220944364.

[81] T. Nakano, T. Yamamoto, "You trust a face like yours," *Humanities and Social Sciences Communications,* vol. 9, no. 1, 2022, doi: 10.1057/s41599-022-01248-8.

[82] J. Newman, *A Taxonomy of Trustworthiness for Artificial Intelligence*, Center for Long-Term Cybersecurity, Berkeley, 2023. [Online]. Available: https://cltc.berkeley.edu/wp-content/uploads/2023/01/Taxonomy_of_AI_Trustworthiness.pdf. [Accessed: Sep. 13, 2023].

[83] European Parliament's Committee on the Internal Market and Consumer Protection and Committee on Civil Liberties, Justice and Home Affairs, *Draft Report on 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*. 2022. [Online]. Available: https://www.europarl.europa.eu/doceo/document/CJ40-AM-732836_EN.pdf. [Accessed: Sep. 13, 2023].

[84] E. Morrow, *Beyond disinformation: deep fakes and false memory implantation*, International Neuroethics Society and International Youth Neuroscience Association, Neuroethics Essay Contest, 2021. [Online]. Available: https://www.dana.org/article/neuroethics-essay-general-audience-2021. [Accessed: Jul. 13, 2023].

[85] Bundesrat, *Beschluss des Bundesrates. Vorschlag für eine Verordnung des Europäischen Parlaments und des Rates zur Festlegung harmonisierter Vorschriften für künstliche Intelligenz (Gesetz über künstliche Intelligenz) und zur Änderung bestimmter Rechtsakte der Union*. 2021. [Online]. Available: https://www.bundesrat.de/SharedDocs/drucksachen/2021/0401-0500/488-21.pdf?__blob=publicationFile&v=1. [Accessed: Sep. 13, 2023].

[86] European Parliament›s Committee on Legal Affairs, *Report on artificial intelligence: questions of interpretation and application of international law in so far as the EU is affected in the areas of civil and military uses and of state authority outside the scope of criminal justice. 2020/2013(INI)*. 2021. [Online] Available: https://www.europarl.europa.eu/doceo/document/A-9-2021-0001_EN.html. [Accessed: Sep. 13, 2023].

[87] R. Michaelson, *Turkish presidential candidate quits race after release of alleged sex tape*, 2023. [Online]. Available: https://www.theguardian.com/world/2023/may/11/muharrem-ince-turkish-presidential-candidate-withdraws-alleged-sex-tape. [Accessed: Sep. 13, 2023].

[88] L. M. Böswald, B. A. Saab, *What a Pixel Can Tell: Text-to-Image Generation and its Disinformation Potential*, Democracy Reporting International, 2022. [Online]. Available: https://democracyreporting.s3.eu-central-1.amazonaws.com/images/6331fc834bcd1.pdf. [Accessed: Sep. 13, 2023].

[89] P. Beuth, A. Demling, M. Hoppenstedt, T. Kleinz, A. Reiner, et al., "Die große Fake-Maschine," *Der Spiegel*, no. 28, pp. 8–15, 2023.

[90] M. Wong, "We Haven't Seen the Worst of Fake News," *The Atlantic*, 2022. [Online]. Available: https://www.theatlantic.com/technology/archive/2022/12/deepfake-synthetic-media-technology-rise-disinformation/672519. [Accessed: Sep. 13, 2023].

[91] S. Kutzner, "Deepfake: Nein, Jordan B. Peterson zog nicht über Baerbock, Lauterbach und Scholz her," Correctiv, 2023. [Online]. Available: https://correctiv.org/faktencheck/2023/03/14/deepfake-nein-jordan-b-peterson-zog-nicht-ueber-baerbocklauterbach-und-scholz-her. [Accessed: Sep. 13, 2023].

[92] D. Neuerer, T. Stiens, *Wie KI zur Gefahr für die Demokratie werden könnte. Handelsblatt*, 2023. [Online]. Available: https://www.handelsblatt.com/politik/deutschland/deepfakes-wie-ki-zur-gefahr-fuer-die-demokratie-werden-koennte/29221078.html. [Accessed: Sep. 13, 2023].

[93] Center on Extremism, *The Dangers of Manipulated Media and Video: Deepfakes and More*, 2023. [Online]. Available: https://www.adl.org/resources/blog/dangers-manipulated-media-and-video-deepfakes-and-more. [Accessed: Sep. 13, 2023].

[94] Der Standard, *Emma Watson liest „Mein Kampf": Trolle feiern Software für Stimmen-Deepfakes*, 2023. [Online]. Available: https://www.derstandard.de/story/2000143117245/emma-watson-liest-mein-kampf-trolle-feiern-software-fuer-stimmen. [Accessed: Sep. 13, 2023].

[95] J. Habgood-Coote, "Deepfakes and the epistemic apocalypse," *Synthese*, vol. 201, no. 3, 2023, doi 10.1007/s11229-023-04097-3.

[96] D. Leufer, *Access Now's submission to the European Commission's adoption consultation on the Artificial Intelligence Act*, Accessed: Sep. 13, 2023. [Online]. Available: https://www.accessnow.org/wp-content/uploads/2021/08/Submission-to-the-European-Commissions-Consultation-on-the-Artificial-Intelligence-Act.pdf. [Accessed: Sep. 13, 2023].

[97] G. Pennycook, T. D. Cannon, D. G. Rand, "Prior exposure increases perceived accuracy of fake news," *Journal of Experimental Psychology: General*, vol. 147, no. 12, pp. 1865–1880, 2018, doi: 10.1037/xge0000465.

[98] M. Karaboga, "Die Regulierung von Deepfakes auf EU-Ebene: Überblick eines Flickenteppichs und Einordnung des Digital Services Act- und KI-Regulierungsvorschlags," in *Digitale Hate Speech. Interdisziplinäre Perspektiven auf Erkennung, Beschreibung und Regulation*, S. Jaki, S. Steger, Ed. Stuttgart: J. B. Metzler, 2023, doi: 10.1007/978-3-662-65964-9_10..

[99] F. Palmiotto, *Detecting Deep Fake Evidence with Artificial Intelligence A Critical Look from a Criminal Law Perspective*, 2023. [Online]. Available: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4384122. [Accessed: Sep. 13, 2023].

[100] Washington State Senate, *Washington State Senate Bill on Defining synthetic media in campaigns for elective office, and providing relief for candidates and campaigns.* S.B. 5152. 2023. [Online]. Available: https://lawfilesext.leg.wa.gov/biennium/2023-24/Pdf/Bills/Senate%20Bills/5152-S.E.pdf?q=20230321103533. [Accessed: Sep. 13, 2023].

[101] Center for an Informed Public, "New WA law requires clear disclosures for 'deepfakes' used in election media," [Online]. Available: https://www.cip.uw.edu/2023/06/09/new-wa-law-deepfake-disclosure-election-media. [Accessed: Sep. 13, 2023].

[102] Bitkom, *Täuschend echt, aber alles Lüge: 63 Prozent haben Angst vor Deepfakes*, 2023. [Online]. Available: https://www.bitkom-research.de/news/taeuschend-echt-aber-alles-luege-63-prozent-haben-angst-vor-deepfakes. [Accessed: Sep. 13, 2023].

[103] J. Laux, S. Wachter, B. Mittelstadt, "Three Pathways for Standardisation and Ethical Disclosure by Default under the European Union Artificial Intelligence Act," *SSRN Electronic Journal*, 2023, doi: 10.2139/ssrn.4365079.

[104] T. Dobber, S. Kruikemeier, F. Votta, N. Helberger, E. P. Goodman, "The effect of traffic light veracity labels on perceptions of political advertising source and message credibility on social media," *Journal of Information Technology & Politics*, 2023, doi: 10.1080/19331681.2023.2224316.

[105] H. Farid, *ChatGPT and Dall-E Should Watermark Their Results*, 2023. [Online]. Available: https://gizmodo.com/chatgpt-dall-e-free-ai-art-should-watermark-results-1850289435. [Accessed: Sep. 13, 2023].

[106] W. Fischer, "California's governor signed new deepfake laws for politics and porn, but experts say they threaten free speech," Business Insider, 2019. [Online]. Available: https://www.businessinsider.com/california-deepfake-laws-politics-porn-free-speech-privacy-experts-2019-10. [Accessed: Sep. 13, 2023].

[107] A. Walorska, *Deepfakes & Disinformation*. Berlin: Friedrich Naumann Fundation, 2020.

[108] J. Langa, "Deepfakes, Real Consequences: Crafting Legislation to Combat Threats Posed by Deepfakes," *Boston University Law Review*, vol. 101, pp. 761–801, 2021. [Online]. Available: https://www.bu.edu/bulawreview/files/2021/04/LANGA.pdf. [Accessed: Sep. 13, 2023].

[109] H. Ajder, J. Glick, *JUST JOKING! Deepfakes, Satire and the Politics of Synthetic Media*, WITNESS, 2021. [Online]. Available: https://cocreationstudio.mit.edu/just-joking/. [Accessed: Sep. 13, 2023].

[110] M. Pawelec, "Deepfakes als Chance für die Demokratie?," in *Digitalisierung und die Zukunft der Demokratie: Beiträge aus der Technikfolgenabschätzung*, A. Bogner, M. Decker, M. Nentwich, C. Scherz, Ed. Baden-Baden: Nomos Verlagsgesellschaft, pp. 89–102, 2022, doi: org/10.5771/9783748928928-89.

[111] F. Gollin, A. Gheorghita, A. Young, O. Ruiz Pilato, X. Chen, *Deepfake. Legal Paper*, Cyber Rights Organization, 2023. [Online]. Available: https://cyberights.org/wp-content/uploads/2023/03/Deepfake-Legal-Paper-CRO2023-1-2.pdf. [Accessed: Sep. 13, 2023].

[112] E. F. Judge, A. M. Korhani, "Deepfakes, Counterfeits, and Personality," *SSRN Electronic Journal,* 2021, doi: 10.2139/ssrn.3893890.

[113] M. Liu, X. Zhang, "Deepfake Technology and Current Legal Status of It," *Proceedings of the 2022 3rd International Conference on Artificial Intelligence and Education (IC-ICAIE 2022)*, 2023, pp. 1308–1314, doi: 10.2991/978-94-6463-040-4_194.

[114] P. Grady, *EU Proposals Will Fail to Curb Nonconsensual Deepfake Porn*, 2023. [Online]. Available: https://datainnovation.org/2023/01/eu-proposals-will-fail-to-curb-nonconsensual-deepfake-porn. [Accessed: Sep. 13, 2023].

[115] N. A. Smuha, "Beyond the Individual: Governing AI's Societal Harm," *Internet Policy Review*, vol. 10, no. 3, 2021, doi: 10.14763/2021.3.1574.

[116] C. Pershan, R. Jonusaite, *User-Guide to the EU Digital Services Act*, EU DisinfoLab, 2022. [Online]. Available: https://www.disinfo.eu/wp-content/uploads/2022/06/20220602_DSAUserGuide_FinalVersion.pdf. [Accessed: Sep. 13, 2023].