DEP *Deportate, esuli, profughe*
Rivista telematica di studi sulla memoria femminile

Deep fake as AI-generated violence against women

di

Sara De Vido[1]

Abstract: This post explains what deep fake is and how it disproportionately affects women. It defines it as a form of "AI-generated gender-based violence against women". The post argues that the current legal framework is weak to respond to the needs of women, and that policies of social platforms are important but insufficient. It will also describe the recent normative evolution at EU level, which, despite being important, fails to capture the gender dimension of the illicit behaviour.

Artificial intelligence (AI) can be exploited to produce deep fake, which consists in the creation of convincing images, audio and video hoaxes. The term describes "both the technology and the resulting bogus content, and is a portmanteau of deep learning and fake." Deep fakes might "invent" art works or produce music, but they might be also used for political misinformation, fraud and for ruining a person's reputation. In this piece, we will specifically focus on non-consensual deep fakes as a form of image-based sexual abuse and AI-generated gender-based violence against women. As a matter of fact, in the majority of cases – studies show (see here and here, for example) – sexual deep fakes are forms of non-consensual pornography. In this analysis, I will first look into the phenomenon of deep fake and its intersectional effects on women and girls, arguing that it is expression of "AI-generated gender-based violence against women." Secondly, I will argue that the current legal framework is weak to respond to the needs of women, and that policies of social platforms are important but insufficient. Thirdly, I will present the recent legal instruments adopted by the European Union (EU) that might provide a legal response, highlighting points of strengths and weakness, and the need for better coordination.

Deep-fake: the phenomenon

The EU Artificial Intelligence Act (EU AI Act), whose final version was approved by unanimity by the Council on 21 May 2024, "deep fake" means AI-generated or manipulated image, audio or video content that resembles existing persons, objects, places or other entities or events and would falsely appear to a person to be authentic or truthful. What the EU AI Act does not recognise is the fact that most deep fakes target women. Platforms regulations do not provide adequate remedy, and the taking down of the images and videos might take too long.

[1] This speech was presented by the author at the Global Conference on AI and Human Rights, held in Ljubljana on 13-14. June 2024.

From a feminist perspective, deep-fake reproduces patterns of discrimination against women in society, diminishes women, trivialises them, uses their bodies as sexual objects, and lead them to silence. It is a violation of privacy, and under this framework non-consensual deep-fake is commonly addressed in domestic legislation, but it affects several human rights and fundamental freedoms: human dignity, sexual expression, and integrity. It causes stress, anxiety, depression. It was argued that deep-fake also produces a collective harm, because there is a risk of normalisation of non-consensual sexual activity and of development of a culture that accepts that what you do in the digital world does not affect reality. Deep-fake can affect all women, especially at the intersection of different grounds of discrimination, and when they play a role in a society as politicians, or bloggers, or writers. Deep-fake is a form of image-based sexual abuse, and it is a form of violence in the digital world or ICT-facilitated violence against women.

ICT-facilitated violence against women (on this concept see also this report) encompasses the different forms of violence committed through computer and communication systems, hardware and software. It includes both online and off-line activities involving any ICT devices, whether connected to networks or not. The forms of violence committed online that qualify as 'gender-based ICT-facilitated violence against women' can be either specific behaviours that are generally executed online and disproportionately negatively affect women and girls (such as non-consensual dissemination of intimate/private/sexual images), or behaviours that are commonly executed offline (and have been defined in that sense in national legislation) but have presented, especially in recent years, an online dimension (e.g. cyberharassment and cyberstalking). Incitement to hatred and violence, otherwise referred to as 'hate speech', is a more specific offence, because the behaviour was originally conceived as a form of violence that required certain forms of dissemination (public dissemination or distribution of tracts, pictures or other material). In recent years, this behaviour has spread and become exacerbated due to the use of ICT and is increasingly targeted at individuals on the basis of their gender, sexual orientation or gender identity. I am arguing here that it is important to distinguish deep-fake from the previous behaviours and stress that deep-fake is a form of "AI-generated gender-based violence against women". ICT not only facilitates deep-fake, it is the technology in itself that might produce violence. Identifying the specificity of AI-generated gender-based violence is also useful because the focus is not only on the dissemination of the material, but also on the creation of the image or the video itself.

The response through "soft" measures

Strengthening the policies of platforms, including reporting mechanisms and swift platform responses are essential, but not enough. Policies and codes of conducts are not compulsory, and can be used as empty label of declared but rarely implemented actions. However, they send a message of commitment to users. In the EU, major online platforms, emerging and specialised platforms, players in the advertising industry, fact-checkers, research and civil society organisations delivered a strengthened Code of Practice on Disinformation following the Commission's Guidance of May 2021. The strengthened Code of Practice on Disinformation was signed on the 16 June 2022, as revision of the 2018 Code. The new Code aims to achieve the objectives of the Commission's Guidance presented in May 2021, by setting a broader range of commitments and measures to counter online disinformation. Commitment No. 14 refers to "malicious deep-fake." Signatories commit to put in place or bolster policies to address both misinformation and disinformation across their services, and to agree on a cross-service understanding of manipulative behaviours, actors and practices not permitted on their services. Among the measures to be adopted, signatories will adopt, reinforce and implement clear policies regarding impermissible manipulative behaviours and practices on their services, based on the latest evidence on the conducts and tactics, techniques and procedures (TTPs) employed by malicious actors, and they will list relevant policies. They also need to report how their respective policies and their implementation address TTPs, threats and harms as well as other relevant threats. In March 2024, the Commission has formally sent requests for information under the Digital Services Act (DSA) to Bing and Google Search (Very Large Online Search Engines, or VLOSEs), as well as to Facebook, Instagram, Snapchat, TikTok, YouTube, and X (Very Large Online Platforms, or VLOPs).

The Commission is requesting these services to provide more information on their respective mitigation measures for risks linked to generative AI, such as so-called 'hallucinations' where AI provides false information, the viral dissemination of deepfakes, as well as the automated manipulation of services that can mislead voters. However, this might not be enough to protect women from violence, because it is a lot based on the discretion of the platform (there is not even a reference to the removal of the material), and the capacity to quickly respond to notifications of violence against women.

The response through "hard law" measures

In the EU, the system of protection in this case comes from three different sources, which will be briefly examined here to make the main argument of the analysis: these are the Digital Services Act, the AI Act and the newly adopted Directive on combating violence against women and domestic violence (VAW Directive). With regard to the DSA, national judicial or administrative authorities can order online platforms to "act against illegal content". What is an illegal content? "illegal content" means any information that, in itself or in relation to an activity, including the sale of products or the provision of services, is not in compliance with Union law or the law of any Member State which is in compliance with Union law, irrespective of the precise subject matter or nature of that law.

Is deep-fake an illegal content? The answer is yes, now, after the adoption of the VAW Directive, which is based on Articles 82(2) and 83(1) Treaty of the Functioning of the EU. The Directive harmonises the elements of some behaviours that fall under the euro-crime of "computer crime," including the non-consensual sharing of intimate or manipulated material. This article was conceived to address the so-called (and inappropriately called) "revenge porn," but the dissemination and the manipulation of material that is sexually explicit are also prohibited. Under Article 5, deep-fake is defined as the intentional conduct of: "producing, manipulating or altering and subsequently making accessible to the public, by means of ICT, images, videos or similar material making it appear as though a person is engaged in sexually explicit activities, without that person's consent, where such conduct is likely to cause serious harm to that person."

Article 23 of the same Directive requires Member States to "take the necessary measures to ensure that online publicly accessible material" as referred in Article 5 is "promptly removed or that access thereto is disabled." Member States shall ensure that the orders and other measures are taken following transparent procedures and are subject to adequate safeguards, in particular to ensure that those orders and other measures are limited to what is necessary and proportionate and that due account is taken of the rights and interests of all relevant parties involved, including their fundamental rights in accordance with the Charter.

On a positive note, the VAW Directive contains the harmonization of the elements of the crime of deep-fake, which means that EU Member States are obliged to transpose this provision into domestic legislation and to guarantee the removal, respecting the DSA and the provisions of the newly-adopted Directive. On the negative side, the fact that one element of the crime is that the conduct is "likely to cause serious harm to that person" fails to capture the inherent harmfulness of these behaviours.

Moving to the EU AI Act, it requires the labelling of deep-fakes as deep-fakes and introduces minimum standards for foundational models, but omits any content moderation. According to the Act, deep-fakes "should also clearly and distinguishably disclose that the content has been artificially created or manipulated by labelling the artificial intelligence output accordingly and disclosing its artificial origin," without affecting 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. There is no reference to the disproportionate impact of AI-generated gender-based violence on women and girls, except for a very short reference in the preamble, where it is written that:

> When improperly designed and used, such systems may be particularly intrusive and may violate the right to education and training as well as the right not to be discriminated against and perpetuate historical patterns of discrimination, for example against women, certain age groups, persons with disabilities, or persons of certain racial or ethnic origins or sexual orientation.

Missing points and conclusions

This is a missing aspect in the EU AI Act, combined with a lack of cross-reference to the VAW Directive.

In conclusion, the EU legal activism has failed to capture the fact that AI-generated material can be a source of empowerment of women and girls, but can also hide misinformation, misogyny and abuse. The example of deep-fake allows us to conclude that there is a strong need for coordination among legal instruments, and, through interpretation, an attempt to bring back to the discourse a neglected gender dimension.

Sara De Vido*, Professoressa associata di diritto internazionale, Università Ca' Foscari Venezia*