Responding to Deepfake non-consensual pornographic images and videos

Public comment for Meta Oversight Board by Dr Afroditi Pina (Reader in Forensic Psychology, University of Kent) and Jenna Harewell (PhD Candidate in Forensic Psychology, University of Kent).

Image based abuse is an umbrella term that encompasses the non-consensual creating, taking or sharing of intimate images, photos or videos and the threats to create, take or share such images (Powell et al., 2019). It also includes unwanted sending of images/videos and pressure/coercion to take and share them. Intimate images now include digitally altered or generated images that are termed as deepfake non-consensual pornography or AI generated image-based abuse (The Image Based Abuse Project, 2024).

Deepfake non-consensual pornographic imagery and videos are causing growing alarm to policy makers and technology platforms around the globe. Deepfakes are not recognized in federal law in the US but there are state-level laws that prohibit it in some forms. The UK Online Safety Act 2023 as of April 16th 2024 specifies that it is an offense to create and distribute private images or videos that are artificially generated in the likeness of someone without their consent. This new offence now stipulates that even if someone creates a sexually explicit deepfake without the intent to share it, but purely want to cause humiliation or distress to the victim they will be committing a criminal offence (Ministry of Justice, 2024). Despite some of the particulars of this new legislation (e.g., the intent to cause harm and distress) having drawn criticism from legal scholars (McGlynn,2024) it is a welcome step in the fight against this newer form of image-based abuse and it also draws more onus on the perpetrator as the mere creation of such material is now criminalized.

As with all aspects of sexual violence and abuse, victims of IBSA are wrongly attributed varying amounts of blame for their victimization and the information they have made available or shared (Pina et al., 2021). Deepfake pornography and the ease with which it can be created (by utilizing just a few pictures of someone's face) has made it clearer to all that for the overwhelming majority of victims, there were never any explicit photos shared, and therefore everyone could be a potential victim.

Although research is providing mixed findings about the gendered nature of image-based abuse as a whole, there is consensus that females report to be impacted to a greater degree (Pina, 2021) and also that pornographic deepfakes are a gendered problem with over 96% of all deepfakes being pornographic (Deeptrace, 2019) and almost all targeting women of all ages; Recently, 11 year old girls were targeted and explicit material shared with their classmates in Spain. The teenagers were fully clothed in the original photos, but an application made them appear realistically naked. Furthermore, many women report being targeted for pornographic deepfakes because of their high visibility occupations as a way to discredit them (e.g., politicians, actors, artists, journalists etc.)

Deepfakes, as all types of sexual violence, have long-lasting deleterious effects on their victims. Victims report feeling violated and dehumanized and their mental health and interpersonal relationships, including their employability, finances and reputations as deeply affected (Henry et al., 2023). Victims often report their memories associated with the affected material and pictures used in the deepfakes to be significantly altered (Guardian, 2023) and the majority experience significant symptomatology of sexual abuse such as PTSD, depression, and anxiety as well as somatic symptoms of insomnia, headaches, palpitations, and gastrointestinal issues. Deepfakes are easy to produce, but not as easy to detect or swiftly remove, trapping their victims in material that is proliferated against their will and oftentimes, depending on jurisdiction, without an option of removal. As it appears

clearly in the two cases that your board is considering, detecting deepfakes is not often easy by the human eye, but also there is no consensus on how to respond to reports on the material and taking it down.

The long-term effects of deepfakes are currently being examined by researchers and scholars across the globe, and new information about their impact is emerging daily. However, we have established that deepfakes can alter perceptions and emotions of not only the victims depicted in them, but the recipients and the people who are exposed to them (Hughes et al., 2023). Hughes et al., found that even if people know that deepfakes exist and that they are being exposed to them, they still form powerful associations and attitudes towards the people depicted in the deepfakes as if they were reality. Deepfakes therefore have the capacity to exploit existing cognitive biases, racism and proliferate misogyny and adversarial gender beliefs.

Research emerging from our lab and specifically the work conducted by Harewell, Pina and Gonidis (under review) is linking psychological and personality factors that are associated with perpetration of image-based abuse (including producing and distributing deepfakes), to be almost identical with those associated with the perpetration of rape and sexual violence. Through a series of studies we have found that myth acceptance surrounding the perpetration and victimization of IBSA, psychopathy, and benevolent sexism are linked with increased proclivity to perpetrate image based abuse. We therefore argue that image-based abuse is part of a continuum of sexual violence and, therefore, should be legally and practically considered as such.

Based on the aforementioned research, we would argue that any deepfake pornographic material that is created in the likeness of someone without their consent should be removed from all technology platforms that it may be shared on and those sharing it should face the appropriate consequences delineated by the platforms and any legal context. Not swiftly removing such material will not only significantly and devastatingly affect those that are targeted in it but will also significantly and long-lastingly affect the perceptions, emotions, attitudes and behaviors of all those exposed to it towards those depicted in this type of material (typically women), thus resulting in further dehumanization and abuse.