AI Deepfake Roundtable 1: Briefing

Glitch and European Network Against Racism (ENAR) are working together to host two workshops on AI harms. This briefing is a summary of the first workshop, with a focus on deepfakes, held on 29th of March 2023. This roundtable brought together a group of anti-racist organisations and deepfake experts (see attendees below) to exchange ideas on how to move research forward in this area. Our goal is to build a community of practitioners and researchers focused on anti-racism and AI harm. In doing so, we will bridge the gap on intersectional approaches to AI harm, particularly deepfake abuse and non-criminal redress.

This briefing contains a summary of the discussion, including:

- Summary of the problem of deepfakes by Henry Ajder, deepfake expert and presenter of BBC's Future Will Be Synthesised
- Addressing systemic and algorithmic racism in the EU by Oyidiya Oji, ENAR
- Glitch's anti-carceral approaches to deepfakes by Gabriela De Oliviera, Glitch
- Areas of focus for moving forward

Please find a recording of the first part of the workshop here: https://youtu.be/hyrY5Usk74c

Attendees

A good number of organisations attended our session. We had the opportunity to share the space with organisations working on digital rights, sex workers rights, anti-racism, and wellbeing platforms.

Summary of the problem: deepfakes (Henry Ajder)

Deepfakes first emerged as a term in late 2017 on reddit as a portmanteau of "deep learning" and fake. It was used to refer to software that swapped women's faces onto pornographic content nonconsensually. Since then, it's expanded to include not just face-swapping but also nudifying apps, cloning voices, and other applications such as lip-synchronisation changes and facial reenactment. These tools are proliferating widely and becoming more sophisticated, fuelled by the rise of generative AI which creates synthetic voice, image and text.

Photoshopping womens' faces has been around for a while. However, there are three reasons why deepfakes are more dangerous than previous forms of photo-editing:

- **Unprecedented realism:** deepfakes are sometimes indistinguishable from real content, can still cause harm even when not fully indistinguishable
- **Efficiency:** when these tools first released, you had to have sophisticated skills to make this. Now you may only need one image, and tools exist to make the workflow easier
- **Accessibility**: tools like Dall-e & Chapt-GPT are gamified and democratised. The fact that these can be accessed in a few clicks, or even embedded in apps like Telegram, has enormous implications.

Understanding the landscape of harm

People are often worried about deepfakes in elections and cybersecurity, but actually on the

ground, 96% of deepfakes were pornographic, 99.9% depicting women (State of Deepfakes report, 2019). This is a global problem, as it is easy to target anyone with a presence online. Accessibility changes not just the amount of content but who is targeted: initially celebrities, but also increasingly private victims. A bot in the web app Telegram allowed people to strip photos of women: 63% said they were doing it to women they knew. Images that were generated could be traced back to social media profiles, and many were underage. One synthetic nudifying website had 38 million hits in two months since its launch, still up and making profits for its owners. These tools ruin lives: digital sexual harassment causes enormous traumatic and reputational harm.

Several open questions remain:

- Ethnographic questions of how this content is created and why?
- What are the racialised impacts of the problem? How does it vary in different regional contexts?
- To better support victims, how do we better put the case to lawmakers in power?

There has been lots of progress in this space in alerting people to the problem of deepfakes, but it is still a challenge getting people to understand that some of the biggest victims of this tech are women.

Addressing systemic and algorithmic racism in the EU (Oyidiya Oji, ENAR):

ENAR works on raising awareness of the racism that many communities are suffering in the EU, linking it historically with systemic harm. AI enhances this historic discrimination. This

means we must bring expertise on digital harms to organisations working on anti-racism and migration. With this technology, we are more exposed to harm than ever. However, this tech can also help us document these harms and prove forms of discrimination we previously were not able to provide proof of. Through this work, we will bring more awareness to our networks around how AI can hurt our communities. In particular, we wish to highlight legislation at the EU level which is relevant to the problem of deepfakes:

Legislation at EU level:

1. EU AI act (proposed in 2021): the framework of the proposed act is a risk based approach. At the beginning, deepfakes were categorised as limited risk, which meant that content had to be tagged as manipulated footage, but there were no penalties for people who didn't tag. This year, in February 2023, there has been an agreement to add deepfakes to a list of high risk models with stricter rules, e.g. risk assessment and human oversight.
2. Directive on Gender-Based Violence (currently being debated): seems more promising but this bill could criminalise non-consensual sharing of perpetrators. However, it only covers explicitly sexual imagery in which people are wholly nude. Furthermore, uf you share it with only a single person, it is considered to not be great harm. Yet we know that in any case in which images are shared online without consent, reputation can be badly damaged.
3. General Data Protection Regulation (effective since 2018): Provides extensive rules for the right to privacy and data protection. In the context of deep fakes, this regulation

refers to data used to train the system - many thousands of pictures have been used to train these models to depict what we want them to depict. In the use of personal data to create deep fakes, consent is needed not just of the people in (output) photo but all the people used for training data. As we have seen in recent cases of the GDPR being used to challenge generative AI models, it can be nearly impossible for companies to prove they have collected this consent. Unfortunately, many of these datasets include images collected nonconsensually or due to sexual abuse (*Content warning: This article includes firsthand accounts of sexual abuse).*

Anti-carceral approaches to deepfakes (Gabriela De Oliveira, Glitch)

As an anti-carceral organisation, we do not support the criminalisation of revenge porn. Instead of investing our energy in criminalisation, we focus on prevention, media literacy, and changing tech companies through tech accountability. The current regulatory landscape is not prepared for AI deepfake abuse, and most countries that are trying to consider deep fakes are doing so through existing legal regimes such as fraud, harassment, and copyright protection in the interim. Countries like Brazil, Nigeria, and Kenya are trying to figure out AI-specific harassment regulation. Here in the UK we are expecting an Online Safety Bill. This currently includes a new criminal offence for deepfakes. Similarly in the US, we've seen banning and criminalisation, usually with a focus on disinformation and misinformation but without attention to the disproportionate way women are targeted. In South Korea, this has been taken more seriously, with a mix between criminalisation and civil fines. In China, there

have been active steps to ban deepfakes, with an approach that mixes criminalisation with more obligations on tech companies, such as risk assessments.

Despite these emerging developments, the regulatory landscape is really far behind technological developments - with this research we aim to get ahead of how this will affect black and racialised women now, before the harm becomes widespread, with clear recommendations to governments and civil society. At the core of this work is understanding harm, and the racialised aspect of deepfake objectification.

Moving forward

Our current research on digital misogynoir and social media platforms, which focuses on text based abuse, shows that Black women are disproportionately affected by the worst kinds of online abuse. AI deepfakes targeting black & racialised women are likely to be increasing. Given research shows pornographic content about black women is disproportionately sexually violent compared to white women[1], it seems likely that black women will be targeted by more sexually violent AI deepfake content and abuse.

In particular, we want to focus future research on the harms experienced by groups like Black people of marginalised genders (like trans women and non-binary people) as well as sex workers, as these groups disproportionately experience fetishisation and dehumanisation.

Discussions in the breakout sessions of the roundtable highlighted overlaps between this

[1] Fritz, N., Malic, V., Paul, B. *et al.* Worse Than Objects: The Depiction of Black Women and Men and Their Sexual Relationship in Pornography. *Gend. Issues* **38**, 100–120 (2021). https://doi.org/10.1007/s12147-020-09255-2

work and the work of other organisations like Chayn and Sister System who work with women, including women who are survivors of sexual exploitation and intimate partner violence. There were also many overlaps with the work of those like Access Now who are working on legal protections and coalition-building in AI harms, and who maintain a focus on racialised women.

We also identified research gaps and areas of priority going forward. As much of the focus has been on women in the public eye, like politicians and celebrities, we need more attention on how this is used in more "everyday" areas like interpersonal violence and abuse against sex workers and other adult performers. We will also explore how to create motivations for tech companies-both large platforms and those who are hosting these websites-and frameworks for them to take responsibility. A big issue going forward will also be the regulation of generative AI, including the data sets they are trained on, their release strategies (fully open sourcing vs more limited API access) and general transparency around their safeguards. Lastly, we want to maintain an anti-carceral focus which looks instead at prevention, public literacy (including on how to spot and respond to deep fakes), and how to address these problems at a community level. In this way, communities that are criminalised and profiled by the police can also access remedies outside of the criminal justice system. With this in mind, we will be hosting a workshop in June with a specific focus on noncriminal redress for AI harms.