AI-ENABLED INTIMATE IMAGE ABUSE

SUMMARY REPORT AND RECOMMENDATIONS

28 JUNE 2022

MINDEROO CENTRE FOR TECHNOLOGY & DEMOCRACY

AI-ENABLED INTIMATE IMAGE ABUSE

SUMMARY REPORT

AI-enabled intimate image abuse is a global phenomenon. The landscape is moving incredibly fast with code-free technology becoming increasingly accessible.

AI-enabled intimate image abuse occurs when harm is caused by the creation and distribution of AI-generated non-consensual intimate images. The definition of 'intimate image' can differ significantly in some communities and does not only refer to sexual imagery.

Today, the average person can use code-free technology with no experience necessary to create images. Through this, we are seeing the democratisation of the ability to cause harm. The technical barriers to entry are now low, and we are seeing synthetically-altered images ("deepfakes") that are increasingly realistic.

On 6 June, the Minderoo Centre for Technology and Democracy organised a workshop to hear from policymakers, industry, and civil society on the topic of AI-enabled intimate image abuse. The session explored what legislation and industry interventions are required to tackle the issue. This workshop followed a scoping session with key stakeholders from academia, industry and civil society in April 2022.

To frame the discussion, it was suggested that the solutions would need to be in policies and the technical specifications of the platforms and services used to spread intimate image abuse.

RECOMMENDATIONS

The following recommendations were discussed:

- **AI-enabled intimate image abuse should be included specifically in Schedule 7 of the Online Safety Bill.** The Online Safety Bill is currently the best opportunity to ensure that all forms of AI-enabled intimate-image abuse are covered by regulation - whether through bringing in existing criminal offences into Schedule 7 or including other forms of abuse in the (as yet not defined) list of harms to adults or codes of practice

- **The definition of online harms in the Online Safety Bill should be more inclusive** (i.e., include reference to gendered-harms in light of the disproportionate extent of online harms faced by women) and take into consideration the interrelation of harms
- **Regulated services should be required to adopt systems and processes that are pre-emptive in detecting and mitigating emerging harms**
- **The conceptualisation of AI-enabled intimate image abuse as an online harm should take into consideration different cultural contexts, especially how it affects women**
- **Regulated services should be compelled to be transparent with their risk assessments**
- **AI-enabled intimate image abuse should be made a criminal offence,** and the offence conceptualised from the viewpoint of the victim, and should not be based perpetrator's motivations in the definition of such an offence
- **The media literacy provision in the Online Safety Bill should be strengthened** to empower users about navigating their safety online

SUMMARY

AI-enabled intimate image abuse workshop hosted by the Minderoo Centre for Technology and Democracy on Zoom (6 June 2022)

Summary by Ann Kristin Glenster, Minderoo Centre for Technology and Democracy Senior Policy Advisor on Technology Governance and Law.

INTRODUCTORY REMARKS

This workshop addressed the need for legal measures to prevent the development of and mitigate harms arising from the use of tools that create AI-enabled intimate image abuse, particularly targeting women. The workshop did not delve into a detailed discussion of what constitutes an AI-enabled intimate image, although some comments were made in this regard. It was particularly noted that it was not clear how the categories of content that are harmful to children and harmful to adults in the Online Safety Bill would address AI-enabled intimate images.

The focus of the discussion was on how regulation should be put in place that would prevent the development of these AI tools in the first place, and how to mitigate against the dissemination of the tools and the images they create. Participants also discussed how regulated services could foresee the risk of harmful content of this kind appearing on their platforms and services, and what mitigation should legally be required in order to prevent the proliferation or targeting of AI-enabled intimate image abuse from taking place.

The workshop began by taking note of how prevalent and easy to use AI tools that enable nudification and intimate image abuse have become. It was highlighted how it has become easy to use these technologies and tools to create images without any technical expertise. As one participant said, this had led to a 'democratisation of the ability to cause harm.' It was also noted that it is becoming increasingly difficult to detect nudified deepfake images at scale as these are becoming increasingly realistic.

There was a particular overarching concern that there is a gender-gap in the Online Safety Bill, and that the conceptualisation of AI-enabled intimate image abuse does not take adequate account of how what constitutes as abuse depends on cultural context.

COLLABORATIVE SOLUTIONS

One of the dominant themes of the workshop were concerns whether the Online Safety Bill would adequately address upstream technologies.

Several participants noted that the proposed regulatory framework was systems-based. In that regard, it was emphasised that the intention was not to wait for the prosecution of someone for harmful content before it was removed, but rather to ensure that regulated services risk assess their services and adopt pre-emptive mitigating systems and processes. However, it was acknowledged by participants that in order to achieve this OFCOM needs to understand the policies that various technology companies already have in place to address harmful content and prioritise user safety. Participants were also concerned that these measures would focus too much on content moderation and not adequately address the systemic issues or design choices that can facilitate emerging harms.

In response, it was explained that regulated services would not be able to rely on existing measures if these were inadequate. Instead, they should be obliged to adopt alternative solutions, and content moderation should only be used to address the small percentage of harm that still slipped through. For example, if a platform can demonstrate an upstream solution which makes it less likely for someone to successfully post content that would be in breach of the regulations – OFCOM should take that into account and potentially be less stringent in terms of punishing individual cases that slip through, because the platform has developed a robust framework.

Participants also considered how far upstream obligations should go through the supply chain, and asked if, for example, third-party filters should fall under the duties imposed on regulated services. It was suggested that the potential for the amplification of harm should be identified in the risk assessments. The workshop participants acknowledged that platforms alone cannot combat the proliferation of tools that enable intimate image abuse – policy interventions would be needed that would require websites to be able to remove content like this from their servers as soon as possible.

It was further iterated that OFCOM would have the powers to audit regulated services to ensure that their policies, and their implementation of these, adequately mitigated harms identified in the risk assessments. It was emphasised that OFCOM's Codes of Practice should include provisions requiring the removal of

harmful content, and that these Codes should recognise how content may have different harmful effects on different communities and especially on women.

It was suggested that in addition to the Online Safety Bill, the UK should adopt a dedicated AI regulatory instrument. It was also suggested that some AI technologies should be banned. For example, AI tools with no other utility than crawling the web for images of women and then to nudify these as realistically as possible should be made illegal.

HARMFUL CONTENT

Some workshop participants noted that the conversation around AI tended to focus on outputs. Some queried whether the Online Safety Bill is too concerned with harm as categories of output, which will not identify the interrelated nature of harm or how the functionality of the regulated services enable harmful content.

Workshop participants also expressed concern that the Online Safety Bill would struggle to keep pace with forms of harm arising from emerging technologies and tools. One participant highlighted that AI-enabled voice abuse was likely to come in the next few years.

Nevertheless, it was noted that the Online Safety Bill should have the flexibility to be extended as emerging harms are identified. As such, there was a general sense through the workshop that AI-enabled intimate images should be added and that this should include cyberflashing.

Some workshop participants also flagged the fact that what constitutes an 'intimate image' and therefore potentially harmful content could be different in different cultural contexts. For example, the use of AI to alter the dress or appearance of someone without their consent, such as removal of usually worn religious dress such as a hijab, could be a form of intimate image abuse.

GLOBAL CONTEXT

Several workshop participants raised the fact that AI-enabled abuse happens in a global context. In that regard, it was noted that the UK regulatory regime had limited territorial reach. There was therefore concern that adopting a UK-only regulatory framework would not be effective. As one workshop participant noted, 'We live in a world of the open web, and we cannot compel these domain providers to take down these sites and tools.'

It was remarked that any legislative intervention to address AI-enabled intimate image abuse should consider what constitutes abuse likely differs around the world, and that in some cases, AI-enabled intimate image abuse had targeted men, especially portraying them as LGBTQ+ in places where this was not culturally acceptable.

REGULATED SERVICES COOPERATION

Some workshop participants noted how the regulatory framework relies on the willingness of regulated services to follow the rules, and that often services and platforms that allow AI-enabled intimate image abuse do not have such intentions. Some messaging services were referenced as examples of platforms that did not care about harmful content (other than child sexual exploitation and abuse (CSEA)). Workshop participants therefore expressed doubts regarding how effective the Online Safety Bill would be in addressing AI-enabled intimate image abuse.

It was also emphasised how regulated services should be required to be transparent regarding how they address harm. One workshop participant queried how the quality of the regulated services' risk assessment and mitigating measures could be verified and trusted.

Some workshop participants raised the point that regulated services commitment to mitigate harm often depended on the gender of their senior management and Boards, and that regulated services that were developed and managed by men may be less likely to prioritise online harm affecting women. It was therefore recommended that OFCOM ensure that online harms facing women are prioritised and mitigating measures are adopted evenly across the industry. Overall, participants were hopeful that the Online Safety Bill, through OFCOM's Codes of

Practice and Guidance, could enable more transparency to ensure that platforms are able to be consistent in their approach rather than be driven by mixed priorities.

CRIMINAL LAW

Several workshop participants were concerned that AI-enabled intimate image abuse is not currently an offence under UK criminal law, although it was recognised that the Law Commission is likely to make recommendations in this regard regarding the law in England and Wales in summer 2022.

Some workshop participants stressed the need to ensure that the regulatory framework was victim-centred, especially in relation to intimate image abuse under criminal law, where legislation continues to focus on the perpetrator's motivation. There was also concern raised regarding support for victims, particularly in contexts where it would be difficult for a victim to explain what had happened. This is particularly the case when the victim is under 18 years of age.

It was suggested that there should be a general principle that a person would need to consent to having their images shared. As such, it was also important to set out how consent should be sought and proved.

Another avenue could be through the criminalisation of the equivalent to intimate image abuse in Scotland, which could under the Online Safety Bill be applied throughout the UK. It was also suggested that the Bill could be amended to require OFCOM to produce a Code of Practice based on the work that is currently being done to address violence against women and girls, which would fall under the Online Safety Bill.

MEDIA LITERACY

There was a call for the provision of media literacy in the Online Safety Bill to be reinstated to an earlier version as there are many people who do not know about the ability of AI to produce nudified deepfake images. Media literacy would make it easier for victims to explain what had happened to them. There also needs to be a strengthening of the teaching of consent in education.

PARTICIPANTS

Participants at the 6 June workshop:

-Damian Collins MP

-Wera Hobhouse MP

-Gina Neff, Minderoo Centre for Technology and Democracy, University of Cambridge

-Henry Ajder, Metaphysic

-Maya Daver-Massion, PUBLIC

-Gabriela De Oliveira, Glitch

-Nima Elmi, Bumble

-Sam Gregory, Witness

-Ann Kristin Glenster, Minderoo Centre for Technology and Democracy, University of Cambridge

-Jeremy Hughes, Minderoo Centre for Technology and Democracy, University of Cambridge

-Clare McGlynn, Durham University

-Katherine Townsend, World Wide Web Foundation

-Raqual Vazquez Llorente, Witness

-Bertie Vidgen, The Alan Turing Institute

-Maeve Walsh, Carnegie UK

-Victoria Wilkinson, Grayling

-Lorna Woods, University of Essex

MINDEROO

CENTRE FOR TECHNOLOGY & DEMOCRACY

Alison Richard Building
7 West Road
Cambridge CB3 9DT

minderoo@crassh.cam.ac.uk

www.mctd.ac.uk