Designing Responsibly To Stop Deepfake Abuse Perpetration

John Twomey

Department of Psychology, University of Limerick, Limerick, Ireland, johntwomeywork@gmail.com

Lero, Science Foundation Ireland Research Centre for Software, Limerick, Ireland

Deepfake abuse involves intentional creation, watching, sharing (or threat thereof) of intimate imagery generated/modified using AI. These generally consist of false images or videos which involve depicting individuals in sexual or explicit contexts without the consent of an individual, most often women. The main question that researchers in this area should be asking is how can we design against deepfake abuse perpetration. Oftentimes, solutions focus overly on technical solutions for recourses and support and not enough on the potentials of working with communities and targeting perpetrators with psychoeducation. We offer a number of potential recommendations and avenues for future research.

CCS CONCEPTS • Psychology • Education • Human computer interaction (HCI)

Additional Keywords and Phrases: Deepfakes, Literacy, Design

1 TECHNOSOLUTIONISM WHEN DESIGNING AGAINST DEEPFAKE ABUSE

With the advent of high quality and accessible non-consensual synthetic intimate imagery (NSII), a number of potential resources and solutions have emerged. Current solutions mainly involve asking victims-survivors (or their representatives) to gather evidence and report abuse materials [17]. While reporting modalities can and do benefit victim-survivors, there are a number of concerns raised about their overall efficacy. Audit work on X has shown that NSII was not removed three weeks after being reported as non-consensual content (when reported as copyright it was removed within two days) [10]. Often existing technological solutions involve removing non-consensual intimate imagery by hashing the imagery and using these hashes to detect and remove them online. However, there are two main concerns with hashing for deepfakes. First, deepfake abuse does not always involve the one image shared multiple times but often consists of a number of different images/videos created [8]. Secondly, there are currently privacy risks around the current encryption of deepfakes and these images can occasionally be reconstructed from the hashes using AI [6]. In a perverse sense, technological solutions for detecting deepfakes may even fuel the improving realism of the technology as they improve the quality of generative adversarial networks which are used to refine the images/videos [3]. The aim of this position is to ask how HCI can contribute to the study of this harm beyond technological solutions. The lens of techno-solutionism is very applicable to deepfake abuse, the attitude that technology can solve the problems which itself creates [1]. Habgood-Coote note that techno-solutionism involves repacking social problems as technological problems [5]. Non-consensual intimate imagery is not a new problem and photoshop was used for this content long before current high-quality generative AI [15]. HCI may benefit from drawing from social approaches to these issues by both targeting how online communities worsen/improve deepfake abuse and how social science research may help understand the harm of the technology and design education or psychoeducational resources.

2 ALTERNATE APPROACHES TO DESIGNING AGAINST DEEPFAKE PERPETRATION

There are three critical issues that need to be considered when designing against deepfake perpetration:

1. Understanding communities and disengagement
2. Implementing deepfake psychoeducation and perpetrator resources
3. Designing to support protective social infrastructures

2.1 Understanding communities and disengagement

A sizable amount of existing literature has focused on the communities of practice which create and share deepfake abuse materials and how this technology as developed on reddit, online forums and GitHub by amateurs who sought to create nonconsensual imagery of celebrity women [4, 12, 14, 18]. While many of these online forums and subreddits have been banned or delisted from search engines, there is still larger concerns around the use of private telegram groups and discord channels dedicated to the creation and sharing of these videos [13]. Reducing access to these may be more challenging and HCI scholars would benefit from reflecting on the existing literature around "exit stories", why people leave polarized online communities [9]. Encouraging individuals to reflect on their membership of these groups and challenging their beliefs may be a successful area for psychoeducation.

2.2 Implementing deepfake psychoeducation and perpetrator resources

Responsibly designing for the use case of deepfake abuse should incorporate psychoeducational materials. Our ongoing work has focused specifically on the role of psychoeducational interventions in reducing the creation and prevalence of deepfake NSII by educating individuals who are at higher risk of being/becoming perpetrators (men between the ages of 18 and 45 who have previously watched online pornography) [16]. The potentials of deepfake psychoeducation involve incorporation into broader interventions targeting technology facilitated gendered violence, teaching school kids about the harms of this technology, and specifically delivering these interventions to men who self-report intentions or past behaviors [16]. How these interventions can be developed to actually reach highly perpetrating individuals is another concern. Perhaps it is worth drawing from existing strategies for other problematic forms of online content. When a person searches results related to child sexual abuse materials, they are already directed towards various different helplines and services (though this implementation varies by country) [2].

2.3 Designing to support protective social infrastructures

As previously stated a large focus of the literature has focused on how the infrastructure which underly deepfake abuse often involve various communities of practice which encourage the creation and sharing of this contents and share values which minimize it's harms [12]. However, outside of increasing our focus on these groups of perpetrators, it may also be useful to consider positive communities of practice which step up to remove this content where technological solutions have failed. These protective communities of practice are often fan communities who protect celebrities from the harms of deepfake abuse, such as during the infamous Taylor Swift deepfake case which saw twitter hashtags relating to the pop star flooded with explicit deepfakes [11]. In response, fan communities sought to flood these hashtags with innocuous content as to reduce the accessibility of offending videos. Similar protective fan behaviors have been noted for fans of k-pop idols [7]. In cases where the design of social media is failing public women, online communities are already responding. Celebrity victim-survivors of deepfake abuse have reported organizing their responses to deepfake abuse and supporting one another through group chats [17]. Supporting these protective community spaces is a vital area for future work.

REFERENCES

[1] Broussard, M. 2018. *Artificial unintelligence: How computers misunderstand the world*. mit Press.

[2] Edwards, G., Christensen, L.S., Rayment-McHugh, S. and Jones, C. 2021. Cyber strategies used to combat child sexual abuse material. *Trends and issues in crime and criminal justice*. 636 (2021), 1–16.

[3] Farid, H. 2022. Creating, using, misusing, and detecting deep fakes. *Journal of Online Trust and Safety*. 1, 4 (2022).

[4] Gamage, D., Ghasiya, P., Bonagiri, V., Whiting, M.E. and Sasahara, K. 2022. Are Deepfakes Concerning? Analyzing Conversations of Deepfakes on Reddit and Exploring Societal Implications. *CHI Conference on Human Factors in Computing Systems* (New Orleans LA USA, Apr. 2022), 1–19.

[5] Habgood-Coote, J. 2023. Deepfakes and the epistemic apocalypse. *Synthese*. 201, 3 (Mar. 2023), 103. https://doi.org/10.1007/s11229-023-04097-3.

[6] Hawkes, S., Weinert, C., Almeida, T. and Mehrnezhad, M. 2024. Perceptual Hash Inversion Attacks on Image-Based Sexual Abuse Removal Tools. *IEEE Security & Privacy*. (2024).

[7] Her, J. and Ringland, K.E. 2025. Mitigating Deepfake Harm in Online Communities: Insights from the BTS ARMY Fandom. (2025), 1–6.

[8] Martin, N. 2021. Image-based sexual abuse and deepfakes: A survivor turned activist's perspective. *The Palgrave Handbook of Gendered Violence and Technology*. (2021), 55–72.

[9] Phadke, S. 2025. Exit Stories: Using Reddit Self-Disclosures to Understand Disengagement from Problematic Communities. *Proceedings of the ACM on Human-Computer Interaction*. 9, 7 (2025), 1–27.

[10] Qiwei, L., Zhang, S., Kasper, A.T., Ashkinaze, J., Eaton, A.A., Schoenebeck, S. and Gilbert, E. 2024. Reporting Non-Consensual Intimate Media: An Audit Study of Deepfakes. *arXiv preprint arXiv:2409.12138*. (2024).

[11] Riedl, M.J. and Newell, A. 2024. Reporting Image-Based Sexual Violence: Deepfakes,# ProtectTaylorSwift, and Platform Responsibility. *Available at SSRN 4919928*. (2024).

[12] Robinson, S., Buckley, J., Ciolfi, L., Linehan, C., McInerney, C., Nuseibeh, B., Twomey, J., Rauf, I. and McCarthy, J. 2024. Infrastructural Justice for Responsible Software Engineering. *Journal of Responsible Technology*. (2024), 100087.

[13] South Korea: The deepfake crisis engulfing hundreds of schools: 2024. *https://www.bbc.com/news/articles/cpdlpj9zn9go*. Accessed: 2024-09-08.

[14] Timmerman, B., Mehta, P., Deb, P., Gallagher, K., Dolan-Gavitt, B., Garg, S. and Greenstadt, R. 2023. Studying the Online Deepfake Community. *Journal of Online Trust and Safety*. 2, 1 (2023).

[15] Twomey, J., Ching, D., Aylett, M.P., Quayle, M., Linehan, C. and Murphy, G. 2024. What Is So Deep About Deepfakes? A Multi-Disciplinary Thematic Analysis of Academic Narratives About Deepfake Technology. *IEEE Transactions on Technology and Society*. (2024).

[16] Twomey, J., Ching, D., Lavoie, E., Geary, A.L., Quayle, M., Linehan, C. and Murphy, G. 2026. Deepfake/Real Harms: An online intervention to reduce deepfake abuse perpetration and myth acceptance. PsyArXiv.

[17] Twomey, J., Foley, S., Robinson, S., Quayle, M., Aylett, M.P., Linehan, C. and Murphy, G. 2025. "What do you expect? You're part of the internet": Analyzing Celebrities' Experiences as Usees of Deepfake Technology. arXiv.

[18] Winter, R. and Salter, A. 2020. DeepFakes: uncovering hardcore open source on GitHub. *Porn Studies*. 7, 4 (2020), 382–397. https://doi.org/10.1080/23268743.2019.1642794.