AI-Generated Image-Based Sexual Abuse: The New Frontier

REBECCA UMBACH, Google, USA
NICOLA HENRY, RMIT, Australia
RENEE SHELBY, Google Research, USA

1 RESEARCHER BACKGROUND

Dr Rebecca Umbach has worked in Trust & Safety Research at Google for almost five years, following a PhD in Criminology and a postdoctoral position in developmental psychology. Her original research area was in non-consensual explicit imagery (NCEI), otherwise known as image-based sexual abuse (IBSA). Her work has helped drive product and policy changes at Google to help victim-survivors report their content more easily and successfully.

Professor Nicola Henry is based at the Social Research Equity Centre at RMIT University. Her socio-legal research, for over 25 years, has focused on the prevalence, nature, and impacts of sexual violence. Since 2013, she has been investigating technology-facilitated sexual violence, including IBSA.

Dr Renee Shelby studies the social impacts of generative AI technologies, with an emphasis on harm-reduction and equity. She received her PhD in sociology of technology from the Georgia Institute of Technology, and has studied technology and gender violence for nine years.

Interest in Generative AI & UGC Rebecca and Nicola's work has contributed foundational knowledge in industry and scholarly settings, including identifying the significant harms faced by victim-survivors, and the relative pervasiveness of this oft-unreported abuse. Existing IBSA prevalence data is primarily limited to Western countries, which we have started to address through our current research agenda. This research is ongoing. We have published our findings on non-consensual sexually explicit deepfakes in 10 countries [14]. We are in the process of writing and submitting the remainder of the papers (on sextortion, victimization, and prevalence); and are simultaneously crafting a survey on artificial intelligence generated—image-based sexual abuse (AI-IBSA) (including "deepfake pornography"), with the aim of providing more color and context into the general public's knowledge about this issue, their beliefs, and related experiences.

2 CONTRIBUTION TO THIS WORKSHOP

We have two main goals in attending this workshop. First, the policy and Trust & Safety teams of many technology platforms around the world are focused on the role user-generated content created through generative AI tools may play in impacting elections, helping propagate misinformation, and scale financial fraud operations. While important work, this narrow focus misses other harmful forms of user-generated content, including IBSA. It is crucial to recognize the significant harms associated with AI-generated IBSA, which at first glance, may feel less critical. Moreover, effective interventions require multi-disciplinary collaboration among ML researchers, engineers, policy teams, and scholarly experts. Second, our goal in attending this workshop is to better understand how other researchers explain complicated concepts to research participants, and identify angles we may have missed in our prospective research plans. Below, we briefly detail the methods and results of our original research, concluding with a brief summary of our prospective survey format.

3 AI-GENERATED IMAGE-BASED SEXUAL ABUSE

Image-based sexual abuse (colloquially known as "revenge porn") is the act of creating, threatening to share, or sharing intimate imagery of someone else without their consent [12]. AI-IBSA is IBSA

Authors' addresses: Rebecca Umbach, Google, USA; Nicola Henry, RMIT, Australia; Renee Shelby, Google Research, USA.

created using generative AI tools, and is malicious user-generated content (UGC) that circulates on social media platforms and through other digital technologies [8]. We place AI-IBSA under the wider umbrella of non-consensual synthetic intimate imagery (NSII), which refers to digitally altered content that is fake but depicts the faces, bodies, and/or voices of real people (and includes cruder creation methods, such as collaging or simple photoshop) [14].

In the online pornography space, dedicated deepfake sites and forums have proliferated. And often, these sites feature famous women [1], including the recent incident with Taylor Swift that led X (Twitter) to temporarily block users from searching for her name on the platform [11]. Less visibly, deepfakes can be weaponized and used for sexual extortion among everyday platform users ("sextortion") [2, 7]. While bad actors previously had to coerce prospective victims to send them real intimate content, now they can extort victims with synthetic content created using generative AI tools.

The harms of IBSA are well-documented [3, 10, 13], including negative impacts on victim-survivors' mental health, career prospects, and willingness to engage with others both online and offline [5, 9]. Potential avenues for reduced harm include (1) legislative interventions, (2) tech company policies around the creation and/or distribution of content,[1] (3) technical tools to help victim-survivors discover and automate takedown requests, and (4) UI treatments on generative AI tools designed to deter users from creating and distributing NSII. A challenge of mitigating AI-IBSA is the proliferation of different companies and tools. Even if some companies implement guardrails, malicious users may circumvent them or turn to less scrupulous tools or open-source software.

4 OUR PRIOR WORK ON AI-IBSA

There is significant media coverage about the risks and harms of deepfake pornography [4, 6, 11], yet no information about the actual prevalence of perpetration and victimization. Moreover, while the majority of content available online appears to be of women [1], these data miss the cases where AI-IBSA is created and not shared widely. The dearth of foundational data makes it challenging to make informed policy decisions or develop design interventions, and limits our ability to assess proposed mitigations. To address this gap, our work [14] examined the following research questions:

RQ1: What is the general public's awareness of, and attitudes towards, AI-IBSA?

RQ2: What is the prevalence of AI-IBSA behaviors (e.g., creating, viewing, and or/sharing images)?

RQ3: How does gender influence both RQ1 and RQ2?

In this study, we surveyed 16,000+ respondents in 10 different countries. The study was quantitative, with optional open-ended questions allowing respondents to add context and additional thoughts.

4.1 Findings on AI-IBSA

Our study generated four main findings about AI-IBSA awareness, behaviors, and gender dynamics. First, despite widespread press coverage, particularly in Western countries, the concept of deepfake pornography is not well-known (less than 30% of all respondents indicated some level of familiarity with the concept of deepfake pornography). Yet on balance, when informed about the concept, respondents thought behaviors associated with non-consensual fake intimate imagery should be criminalized. Second, self-reported victimization prevalence was relatively low, as compared to more traditional image-based sexual abuse prevalence rates. Third, the most common behaviors self-reported by respondents were passive and readily available (e.g., the consumption of celebrity deepfake pornography, reported by 6% of respondents), with behaviors requiring effort and/or money being very rare (e.g., creating deepfake pornography). Finally, compared to women, men

[1] e.g., Reddit's de-platforming of subreddits dedicated to deepfake pornography or Google's policy for removal of NSII.

think of these behaviors as less bad, and are also more likely to report being both victimized by and perpetrating AI-IBSA.

As practitioners and designers think about policy, design, or technical interventions, it may be helpful to break this issue down into relevant behaviors and consider intervention options that should be considered complementary, as opposed to solely sufficient, at each point in the process. For example, the increased ease of access to NSII creation technologies was the catalyst for this project, and represents the first place at which interventions could be made. Are there ways, for example, to limit these tools' abilities to create nude or intimate imagery? If creators get to the point where barriers have been hacked, are there warning messages that can either highlight associated harms to targets, or potential punishments associated with the creation or distribution of such content? After creation, are there ways to build in indicators of provenance (e.g., watermarking) that signal the synthetic nature of the content? Finally, assuming most impediments can be circumvented by a motivated creator, empathetic and multi-modal resources for victims could include both tooling to discover where relevant content is published, and help for seeking take downs and recourse more broadly. These are primarily design interventions, but overarching all of these would be the societal interventions of educational programming, and relevant legislation making clear potential punishments associated with each type of behavior (extortion, posting/distribution, etc), paired with prosecutorial actions that indicate a willingness to hold perpetrators to account.

5 PROSPECTIVE RESEARCH ON AI-IBSA

While our existing work was derived from a larger survey on image-based sexual abuse, our prospective survey focuses exclusively on AI-generated IBSA. We plan to ask similar prevalence and attitudinal questions, as well as questions around distribution, motivations, and consumption. We are still navigating concerns about the rapidly evolving landscape of generative AI tools, level of sophistication and knowledge of survey respondents, and ethical concerns.

Our hope is that this workshop will, in addition to being an opportunity to meet other researchers in this issue area, be a space for ideation around:

(1) How to accurately study the public's knowledge, attitudes, and beliefs about generative AI using survey methods.
(2) Mitigation strategies others have already attempted for the aforementioned concerns about malicious UGC.
(3) Specific at-risk populations to consider.

REFERENCES

[1] Henry Adjer, Giorgio Patrini, Francesco Cavalli, and Laurence Cullen. 2019. *The State of Deepfakes: Landscape, Threats, and Impact.* Technical Report. Deeptrace Labs. https://regmedia.co.uk/2019/10/08/deepfake_report.pdf.

[2] Nadisha-Marie Aliman, Leon Kester, and Roman Yampolskiy. 2021. Transdisciplinary AI observatory—retrospective analyses and future-oriented contradistinctions. *Philosophies* 6, 1 (Jan. 2021), 6. https://doi.org/10.3390/philosophies6010006

[3] Samantha Bates. 2017. Revenge porn and mental health: A qualitative analysis of the mental health effects of revenge porn on female survivors. *Feminist Criminology* 12, 1 (2017), 22–42.

[4] Matt Burgess. 2023. *Deepfake Porn Is Out of Control.* Wired. https://www.wired.com/story/deepfake-porn-is-out-of-control/

[5] Danielle Keats Citron and Mary Anne Franks. 2014. Criminalizing revenge porn. *Wake Forest Law Review* 49 (May 2014), 345–391.

[6] Samantha Cole. 2018. *We Are Truly Fucked: Everyone Is Making AI-Generated Fake Porn Now.* Vice. https://www.vice.com/en/article/bjye8a/reddit-fake-porn-app-daisy-ridley?__twitter_impression=true

[7] Hany Farid. 2022. Creating, using, misusing, and detecting deep fakes. *Journal of Online Trust and Safety* 1, 4 (Sept. 2022), 1–33. https://doi.org/10.54501/jots.v1i4.56

[8] Nicola Henry and Asher Flynn. 2019. Image-based sexual abuse: Online distribution channels and illicit communities of support. *Violence Against Women* 25, 16 (2019), 1932–1955.

[9] Nicola Henry, Clare McGlynn, Asher Flynn, Kelly Johnson, Anastasia Powell, and Adrian J Scott. 2020. *Image-based sexual abuse: A study on the causes and consequences of non-consensual nude or sexual imagery.* Routledge, New York, NY.

[10] Mudasir Kamal and William J Newman. 2016. Revenge pornography: Mental health implications and related legislation. *Journal of the American Academy of Psychiatry and the Law Online* 44, 3 (2016), 359–367.

[11] Luba Kassova. 2023. *Tech bros need to realise deepfake porn ruins lives – and the law has to catch up.* The Guardian. https://www.theguardian.com/global-development/2024/mar/01/tech-bros-nonconsensual-sexual-deepfakes-videos-porn-law-taylor-swift

[12] Clare McGlynn, Erika Rackley, and Ruth Houghton. 2017. Beyond "revenge porn": The continuum of image-based sexual abuse. *Feminist Legal Studies* 25 (2017), 25–46. https://doi.org/10.1007/s10691-017-9343-2

[13] Yanet Ruvalcaba and Asia A Eaton. 2020. Nonconsensual pornography among US adults: a sexual scripts framework on victimization, perpetration, and health correlates for women and men. *Psychology of Violence* 10, 1 (2020), 68.

[14] Rebecca Umbach, Nicola Henry, Gemma Beard, and Colleen Berryessa. 2024. Non-Consensual Synthetic Intimate Imagery: Prevalence, Attitudes, and Knowledge in 10 Countries. arXiv:2402.01721 [cs.CY]