Measuring and mitigating harms at the intersections of **generative AI & image-based sexual abuse**

Natalie Grace Brigham, ***University of Washington***

SUPA 2024

Content warning: This talk discusses image-based sexual abuse

About me

- Recent BS/MS in computer science graduate from **University of Washington**
  - Advised by **Tadayoshi Kohno**
- Interested in studying **sociotechnical harms of AI**
- Presenting **“‘Violation of my body:’ Perceptions of AI-generated non-consensual (intimate) imagery”** during Tuesday’s technical sessions
- Fun fact: **First paper/conference/workshop!**

“deepfake”

deep learning + fake

a bit of **etymology**

- Term coined in **December 2017** by Reddit user who created **r/deepfakes subreddit**
- Subreddit popularized as source of **synthetic, non-consensual intimate imagery**
- In **February 2018**, Reddit (and others) **banned or removed** some deepfake communities and content
- Non-consensual, synthetic sexual content **continues to be produced and circulated** on other forums (e.g., MrDeepFakes)

a bit of **terminology**

“deepfake” → **“synthetic media”**

It has been suggested that this article be merged with *Synthetic media*. (Discuss) *Proposed since January 2024.*

a little more **terminology**

"deepfake porn"

→ **"AI-generated non-consensual intimate imagery"**

AIG - NCII

→ "Non-consensual synthetic intimate imagery"

→ "Non-consensual synthetic explicit imagery"

AIG-NCII Prevalence

The State of Deepfakes - Deeptrace Labs (2019*)

- Out of 14,678 synthetic videos identified online, **98% were “pornographic”**
- Found that sexually explicit videos **exclusively targeted women**
- **134+ Million views** across top four “deepfake pornography” websites

**Measurements taken in December 2018*

Non-Consensual Synthetic Intimate Imagery: Prevalence, Attitudes, and Knowledge in 10 Countries - Umbach, Henry, Beard & Berryessa (2024*)

- Of 16,000 respondents, **2.2% indicated personal victimization** and **1.8% indicated perpetration behaviors**
- **Men reported significant higher rates of victimization and perpetration** experiences related top creation and threats

**Measurements taken in mid-2023*

Attitudes Towards AIG-NCII

- People favored **criminalization** of both sexual and non-sexual deepfakes *(UK-based)*[1]
- Deepfakes of **celebrities** were perceived as **less criminal/harmful** *(US-based)*[2]
- Across 10 countries, average **awareness of AIG-NCII was low**, but respondents believed victims had a **right to be upset**[3]
- People were far more accepting of someone seeking out AIG-NCII than creating or sharing it *(US-based)*[4]

(1) Matthew B. Kugler and Carly Pace. Deepfake Privacy: Attitudes and Regulation. *Nw. UL Rev.*, 116:611, 2021.

(2) Dean Fido, Jaya Rao, and Craig A. Harper. Celebrity status, sex, and variation in psychopathy predicts judgements of and proclivity to generate and distribute deep- fake pornography. *Computers in Human Behavior*, 129:107141, 2022.

(3) Rebecca Umbach, Nicola Henry, Gemma Faye Beard, and Colleen M. Berryessa. Non-consensual synthetic intimate imagery: Prevalence, attitudes, and knowledge in 10 countries. In Proceedings of the CHI Conference on Human Factors in Computing Systems, CHI '24, New York, NY, USA, 2024. Association for Computing Machinery.

(4) Natalie Grace Brigham, Miranda Wei, Tadayoshi Kohno, and Elissa M. Redmiles. "Violation of my body:" Perceptions of AI-generated non-consensual (intimate) imagery. 2024. Available online at https: //arxiv.org/abs/2406.05520

Legal & policy-based work

- Advocacy for IBSA victim-survivors

- Legal scholarship around IBSA and AIG-NCII (e.g., Danielle Citron's work)
- White House executive order on AI & call to action on IBSA
- Various state legislative actions (some explored in this AP article)

Looking Forward

Future research

Further measurements of

- **AIG-NCII online**
- Creation and sharing **ecosystem**
- **Attitudes** (over time and across countries)

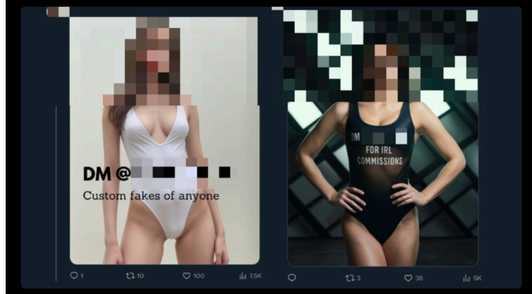
Technical interventions:

- Designing systems for **consent**
- **Detecting** AIG-NCII
- **Preventing** generation (e.g., red teaming approaches)

Source:  Emanuel Maiberg @ ~~404~~

An evolving threat

“Nudify” apps and services

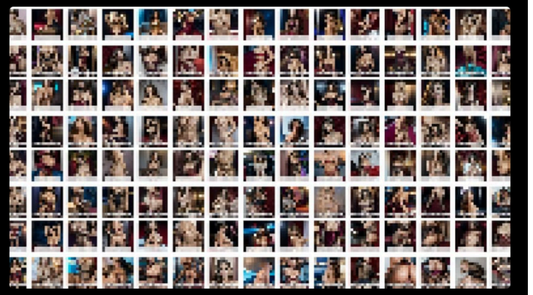


NEWS

‘IRL Fakes:’ Where People Pay for AI-Generated Porn of Normal People

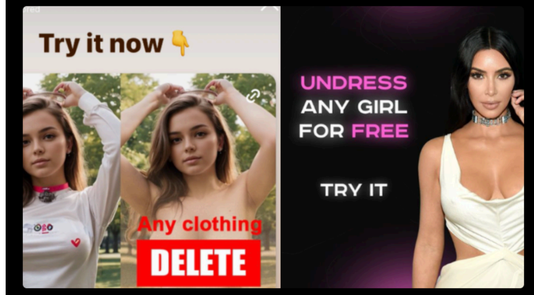
NEWS

Pornhub’s Biggest Star Is Promoting an AI Nonconsensual ’Nudify’ App



NEWS

How Makers of Nonconsensual AI Porn Make a Living on Patreon



AI

Instagram Advertises Nonconsensual AI Nude Apps

NEWS

Google Search Includes Paid Promotion of “Nudify” Apps

NEWS

This Camera Turns Every Photo Into a Nude

New jailbreaks

NEWS

Microsoft Closes Loophole That Created AI Porn of Taylor Swift

NEWS

Users ‘Jailbreak’ AI Video Generator to Make Porn

Can ethical, pornographic synthetic media exist?

What training data could be used?

How do we define consent in this context?

What guardrails could be put in place? How?

"We're exploring whether we can responsibly provide the ability to **generate NSFW content** in age-appropriate contexts … We look forward to better understanding user and **societal expectations of model behavior** in this area."

Model Spec Document [1]

"We want to ensure that people have **maximum control** to the extent that it doesn't **violate the law or other peoples' rights**, but enabling **deepfakes is out of the question**, period."

"Depends on your **definition of porn**…These are the exact conversations we want to have."

Joanne Jang, Model Lead [2]

Can ethical, pornographic synthetic media exist?

What training data could be used?

How do we define consent in this context?

What guardrails could be put in place? How?

1 https://cdn.openai.com/spec/model-spec-2024-05-08.html#dont-respond-with-nsfw-content
2 https://www.npr.org/2024/05/08/1250073041/chatgpt-openai-ai-erotica-porn-nsfw

Measuring and mitigating harms at the intersections of **generative AI & image-based sexual abuse**

Natalie Grace Brigham
University of Washington
nbrigham@uw.edu

SUPA 2024

Miranda will be presenting "Understanding Help-Seeking and Help-Giving on Social Media for Image-Based Sexual Abuse" at USENIX on Thursday!

I'll be presenting "'Violation of my body:' Perceptions of AI-generated non-consensual (intimate) imagery at SOUPS on Tuesday!

Appendix

Image-Based Sexual Abuse (IBSA)

The non-consensual creation, distribution, or threats made with intimate images

- Various types (e.g., sextortion, AIG-NCII, pressurized sexting, cyberflashing, nonconsensual explicit imagery)[1]
- Victim-survivors can experience:
  - Health consequences (e.g., post-traumatic stress disorder, anxiety, depression, and greater somatic burdens)[2,3,4,5]
  - Social harms (e.g., isolation, lowered self-esteem, trust issues, and unhealthy coping mechanisms)[2,6]
  - Victim-blaming attitudes when seeking support or justice after IBSA[7]

Miranda will be presenting "Understanding Help-Seeking and Help-Giving on Social Media for Image-Based Sexual Abuse." at USENIX Security on Thursday!

(1) Miranda Wei, Sunny Consolvo, Patrick Gage Kelley, Tadayoshi Kohno, Tara Matthews, Sarah Meiklejohn, Franziska Roesner, Renee Shelby, Kurt Thomas and Rebecca Umbach. Understanding Help-Seeking and Help-Giving on Social Media for Image-Based Sexual Abuse. *USENIX Security, 2024.*

(2) Samantha Bates. Revenge Porn and Mental Health: A Qualitative Analysis of the Mental Health Effects of Revenge Porn on Female Survivors. *Feminist Criminology*, 12(1):22–42, January 2017.

(3) Asia Eaton, Holly Jacobs, and Yanet Ruvalcaba. Nationwide Online Study of Nonconsensual Porn Victimization and Perpetration. Technical report, Cyber Civil Rights Initiative, 2017.

(4) Antoinette Huber. 'A shadow of me old self': The impact of image-based sexual abuse in a digital society. *International Review of Victimology*, 29(2):199–216, 2023.

(5) Yanet Ruvalcaba and Asia A Eaton. Nonconsensual Pornography among US Adults: A Sexual Scripts Frame- work on Victimization, Perpetration, and Health Correlates for Women and Men. *Psychology of Violence* 10(1):68, 2020.

(6) Clare McGlynn, Kelly Johnson, Erika Rackley, Nicola Henry, Nicola Gavey, Asher Flynn, and Anastasia Powell. 'It's Torture for the Soul': The Harms of Image-Based Sexual Abuse. *Social & Legal Studies*, 30(4):541– 562, 2021.

(7) Asher Flynn, Elena Cama, Anastasia Powell, and Adrian J Scott. Victim-blaming and image-based sexual abuse. *Journal of Criminology*, 56(1):7–25, 2023.

“Deepfake” communities’ attitudes

- Pro-deepfake attitudes among **Reddit** users, who supported building a marketplace for such content **regardless of the consequences**[1]
- Widder et al. found positive attitudes towards deepfakes but also **heightened misuse concern** in **open-source** tool’s community[2]
- **MrDeepFakes**, the self proclaimed largest platform for AIG-NCII, hosts a community on its forum[3]
  - Sharing is limited to content depicting celebrities
  - Doubles as hub for purely **technical advice**
  - **Defensive** culture with distrust for other spaces (e.g., Reddit)

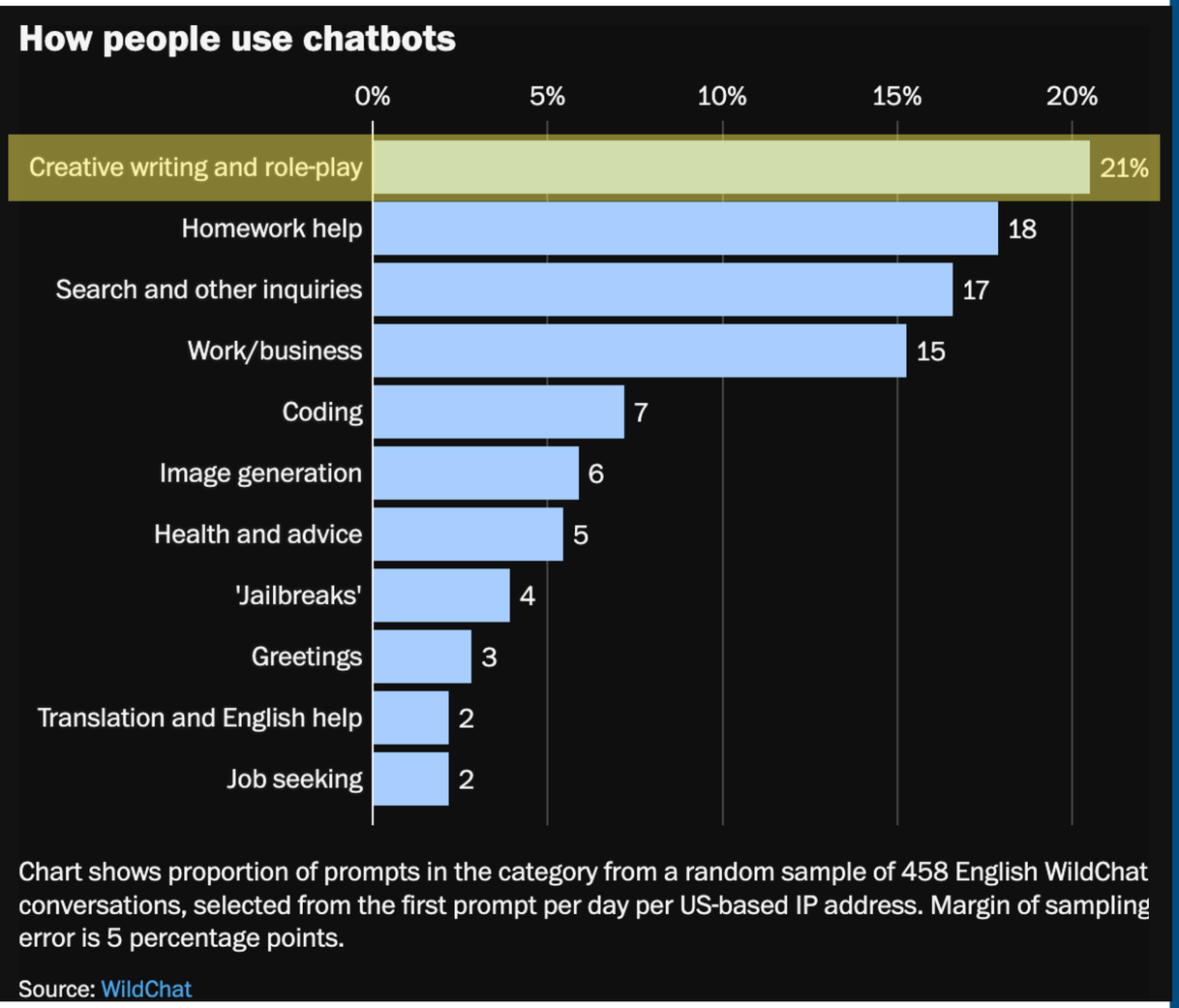
(1) Dilrukshi Gamage, Piyush Ghasiya, Vamshi Bonagiri, Mark E. Whiting, and Kazutoshi Sasahara. Are Deepfakes Concerning? Analyzing Conversations of Deepfakes on Reddit and Exploring Societal Implications. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI ’22, pages 103:1–103:19, New York, NY, USA, 2022. Association for Computing Machinery.

(2) David Gray Widder, Dawn Nafus, Laura Dabbish, and James Herbsleb. Limits and Possibilities for “Ethical AI” in Open Source: A Study of Deepfakes. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, FAccT ’22, page 2035–2046, New York, NY, USA, 2022. Association for Computing Machinery.

(3) Brian Timmerman, Pulak Mehta, Progga Deb, Kevin Gallagher, Brendan Dolan-Gavitt, Siddharth Garg, and Rachel Greenstadt. Studying the Online Deepfake Community. *Journal of Online Trust and Safety*, 2(1), Sep. 2023.

What do people really ask chatbots? It's a lot of sex and homework.

The Washington Post



https://www.washingtonpost.com/technology/2024/08/04/chatgpt-use-real-ai-chatbot-conversations/

Can ethical, pornographic synthetic media exist?

What training data should be used?

What would define consent and likeness in this context?

What guardrails should be put in place? How?