DEFENSE-IN-DEPTH AGAINST AI GENERATION OF NCII & CSAM

Georgetown University — Dr. Elissa M. Redmiles, Dr. Lucy Qin, Dr. Sarah A. Bargal

EPFL École Polytechnique Fédérale de Lausanne — Dr. Ana-Maria Cretu, Klim Kireev, Dr. Carmela Troncoso

University of Washington — Natalie G. Brigham, Dr. Miranda Wei, Dr. Tadayoshi Kohno

Boston University — Princessa Cintaqia, Dr. Allison McDonald

UC San Diego — Arshia Arya, Dr. Deepak Kumar

FIU Florida International University — Dr. Asia Eaton

There is no technical "silver bullet" to stop the creation of **S**ynthetic **N**on-**C**oncensual **I**ntimate **I**magery.

We take a **defense-in-depth** approach: strengthening technical infrastructure & social norms to reduce harm.

SHAPE SOCIAL NORMS: TECH-ENABLED DETERRENCE MESSAGING

Our research finds that
90%+ of Americans
think creating & sharing SNCII is unacceptable

But barely
50%
think viewing is unacceptable

We're combining principles from psychology & persuasion with computational analysis to develop & target deterence messages against SNCII creator & viewer searches for tools & content.

We will evaluate our messages in the field with two large platforms already on board as partners.

Brigham, N. G., Wei, M., Kohno, T., & Redmiles, E. M. "Violation of my body:" Perceptions of AI-generated non-consensual intimate imagery. In Twentieth Symposium on Usable Privacy and Security (SOUPS 2024) (pp. 373-392).

RIGOROUS TECHNICAL EVALUATIONS FOR EFFECTIVE PROTECTION AGAINST CSAM

One proposal to stop generative models from producing CSAM and NCII is concept cleaning, the rigorous filtering training data related to a harmful concept (e.g., child, nudity) to prevent harmful content generation.

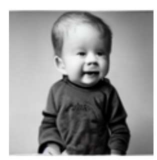
We implemented this proposal by retraining generative models from scratch. Preliminary results (right) show that after automated filtering, we can still produce images of a CSAM proxy concept: a child wearing glasses.

Our ongoing assessment focuses on measuring the impact of filtering on difficulty: odds of generation across threat models.

Even a model trained on a dataset filtered to remove images of children can still generate an image of a child.

Output of text2image model prompted with: "picture of a child" and trained on...

unfiltered data



filtered data ~94% of images of children removed from training data

FORECAST NEW ATTACKS

"We really want to get to a place where we can enable NSFW stuff for your personal use in most cases… **but not do stuff like make deepfakes.**"

– Sam Altman, OpenAI

We are studying the risk that someone can (un)intentionally create SNCII using only text. Using prompt inversion, we upper bound the odds that a mode produces SNCII when prompted with a detailed text description of a mental image (of a fantasy, or a victim), without naming any living individual or providing an initial image depicting them.

CURATE ETHICAL BENCHMARK DATASETS

To build AI-based content moderation tools, developers need datasets of nude content:

1. Many if not all open-source & academic datasets lack the consent of the nude image subjects depicted
2. Datasets must be curated with the adult subject's consent + participatory governance to avoid perpetuating the same harms content moderation tools aim to prevent

Cintaqia, P., Arya, A., Redmiles, E.M., Kumar, D., McDonald, A., Qin, L. Abuse in the Name of Safety: Stop the Non-consensual Use of Intimate Images in Research. Forthcoming in NeurIPS 2025 (Position Paper) and AIES 2025.

elissa.redmiles@georgetown.edu

NUDIFICATION TOOLS HARM CHILDREN & ADULTS
FOR AS LITTLE AS 6 CENTS AN IMAGE

UNIVERSITY of FLORIDA — Kevin Butler, Cassidy Gibson, Patrick Traynor

Georgetown University — Elissa M. Redmiles

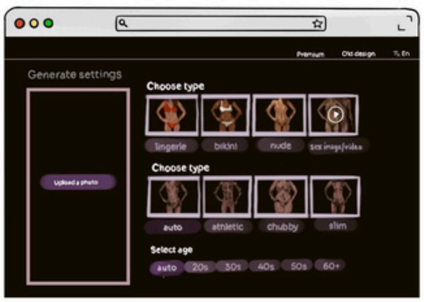
UNIVERSITY of WASHINGTON — Natalie G. Brigham, Tadayoshi Kohno

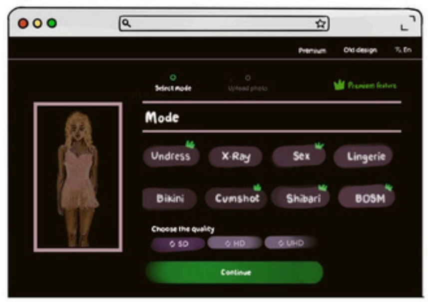
"You don't dare to talk to the person you love? [No problem]! Upload her photo her and find out what she looks like when she...(haha) 👀 😏 😈

Analyzing AI nudification tools exposes their true capabilities and **the specific threats they pose to victims.**

C. GIBSON, D. OLSZEWSKI, N. G. BRIGHAM, A. CROWDER, K. R. B. BUTLER, P. TRAYNOR, E. M. REDMILES, AND T. KOHNO, "ANALYZING THE AI NUDIFICATION APPLICATION ECOSYSTEM," IN PROCEEDINGS OF THE 34TH USENIX SECURITY SYMPOSIUM (USENIX SECURITY '25), USENIX ASSOCIATION, 2025.

OUR PEER-REVIEWED RESEARCH ANALYZES HOW 20 NUDIFICATION TOOLS OPERATE





Lack of Safe Guards

- Not a single one of the twenty platforms includes any explicit consent check to verify the person in the image gave consent to be nudified.
- Only 7 out of 20 tools analyzed have terms of service that require the image subject to be over 18. None validate age.
- Discussion boards for some tools are a trading hub for CSAM and contain guides to generate these images.

Low Cost

- Free options for users willing to join queues and watch ads for nude image generation.
- Paid options range from $0.06 to $0.72 for an image without a queue, averaging $0.34.
- Similar edits by humans cost $20+, charge for additional edits, and can take weeks.

Features

- Nudify applications don't just nudify: over half of the studied applications could put image subjects into sexual scenes, including videos.
- Many purported to be able to change the victim's age, body size, etc.
- Nudify applications are built to spread: many of the websites we study offer affiliate marketing & referral programs and some even offer API access & business manager services others can use to spin up their own nudify platforms

WE IDENTIFY STAKEHOLDERS THAT CAN DISRUPT THESE TOOLS TO STOP VICTIMIZATION

We identify critical stakeholders in the AI nudification ecosystem—developers, platforms, and financial intermediaries—that can disrupt these tools before they reach victims. This approach shifts the burden away from survivors and focuses instead on cutting off the production, monetization, and distribution pathways that enable these tools to thrive.

1 Single Sign On

Single Sign On Providers give websites some level of "legitimacy". 12/20 tools use single sign on providers such as:

- 12/20 Google
- 4/20 Discord
- 3/20 iCloud
- 1/20 X

2 Payment Processors

Payments Processors help decrease barrier of entry for buyers - pressure on this third party has shown to be beneficial in the past.

- 16 allow for Cryptocurrency Gateways, use in sexual abuse can threaten the legitimacy of cryptocurrency
- 8 use Paypal

3 Affiliate Marketing

Many AI Nudification tools use affiliate marketing to advertise.

- 9/20 tools use affiliate marketing
- Offer 20%-50% shared revenue

4 Web Hosting Agencies

While legitimate websites use web hosting agencies to host their websites, these applications do as well:

- 12/20 use Cloudflare
- 7/20 use Amazon Web Services

5 Source Models

- Many of these applications are trained on top of open-source models or use computer vision techniques. The long history of open-sourcing models with the functionality needed for these AI Nudification websites means developers do not always need extensive expertise to run these applications. There is emerging discussion in the ML research community on whether models with risky capabilities should be access restricted.

cleighgibson@gmail.com

elissa.redmiles@georgetown.edu

safe digital intimacy.org Innovating safe digital intimacy through tech research and design

PRISM Center for Privacy and Security of Marginalized and Vulnerable Populations

NSF National Science Foundation