

Tattle

# What Can We Do About Open Weight Models Enabling NCSI?

Tarunima Prabhakar
Tattle Civic Technologies

# AI Developments Are Motivated by Sexual Content Production

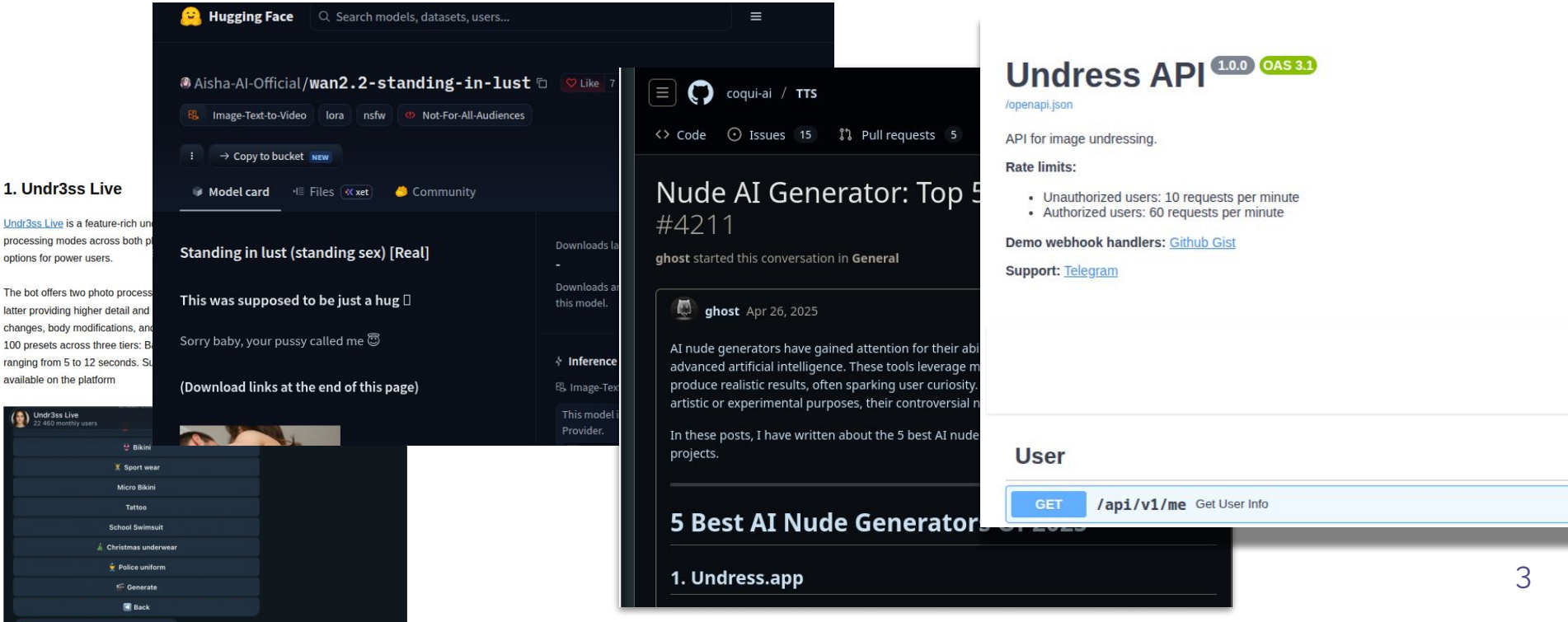
- 98% of ‘deepfakes’ are pornographic. Of which 99% are of women. -Security Hero, 2023
- The term ‘deepfake’ comes from a Reddit thread devoted to porn
- FaceSwaps (GAN) -> AI generated image/ videos (Diffusion models)



Portrait of a Young Woman by Sandro Botticelli (1480) on the left, "nudified" through deepfake technology on the right, with censorship added afterward
Source: Wiki Commons

Tattle

# AI Generated Sexual Content is Everywhere



# The Harm Is Real

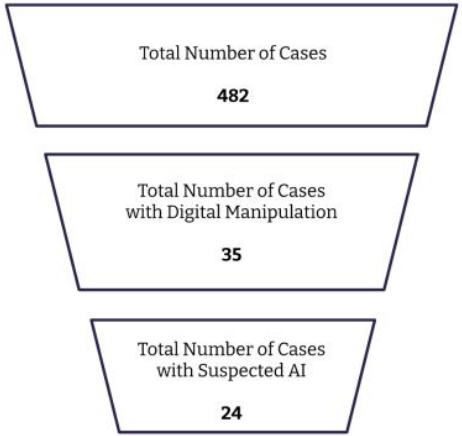
The survivor…felt deeply shamed and socially marked, as though she had been "involved in something dirty" and felt guilty that she was involved in this.

- Make it Real Report

Tattle

## Evidence from Meri Trustline

Since 2022, over 482 survivors have sought support from Meri Trustline. The majority of cases involve some digital content. This could be sharing of NCII on social media or in messaging groups, or recording a person viewing sexual content for extortion.[38] 35 cases reported have involved digitally manipulated content. The spike in digitally manipulated content is correlated with advancements in AI. Most of the content (24 out of 35) reported as digitally manipulated is suspected to involve some AI-based manipulation.



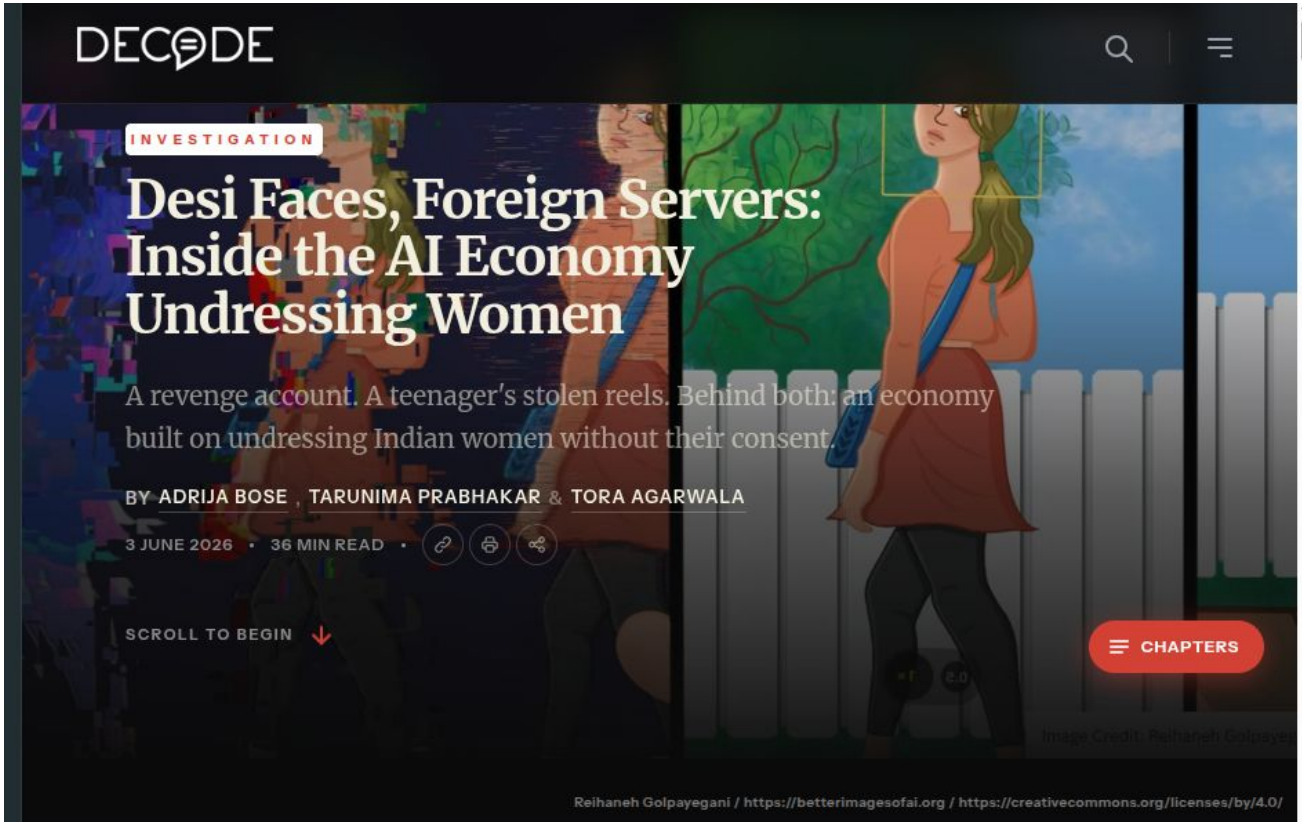
# The Issue is Consent

The problem is not sexual content but the non-consensual generation of sexual content.

AND

Consent cannot be technologically verified.

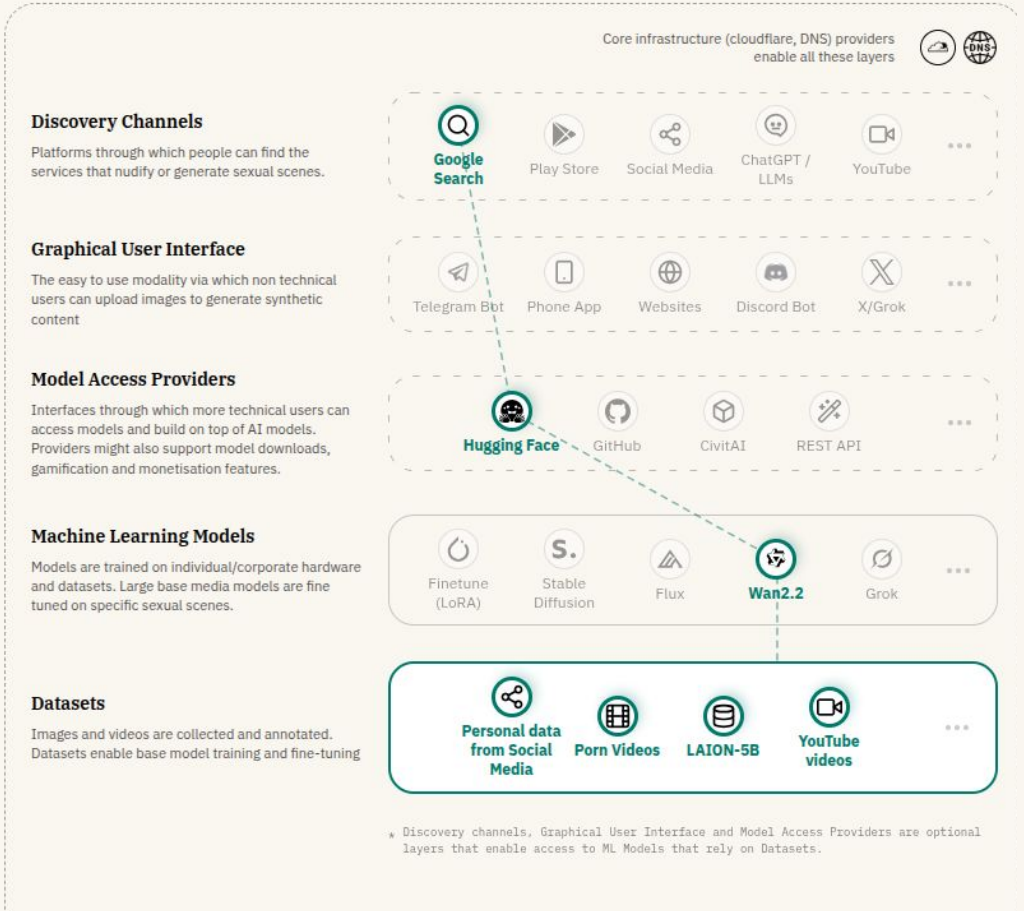
Tattle


DECODE
INVESTIGATION
Desi Faces, Foreign Servers: Inside the AI Economy Undressing Women
A revenge account. A teenager's stolen reels. Behind both: an economy built on undressing Indian women without their consent.
BY ADRIJA BOSE , TARUNIMA PRABHAKAR & TORA AGARWALA
3 JUNE 2026 • 36 MIN READ •
SCROLL TO BEGIN
CHAPTERS
Reihaneh Golpayegani / https://betterimagesofai.org / https://creativecommons.org/licenses/by/4.0/

Tattle

## Infrastructure Tech Stack

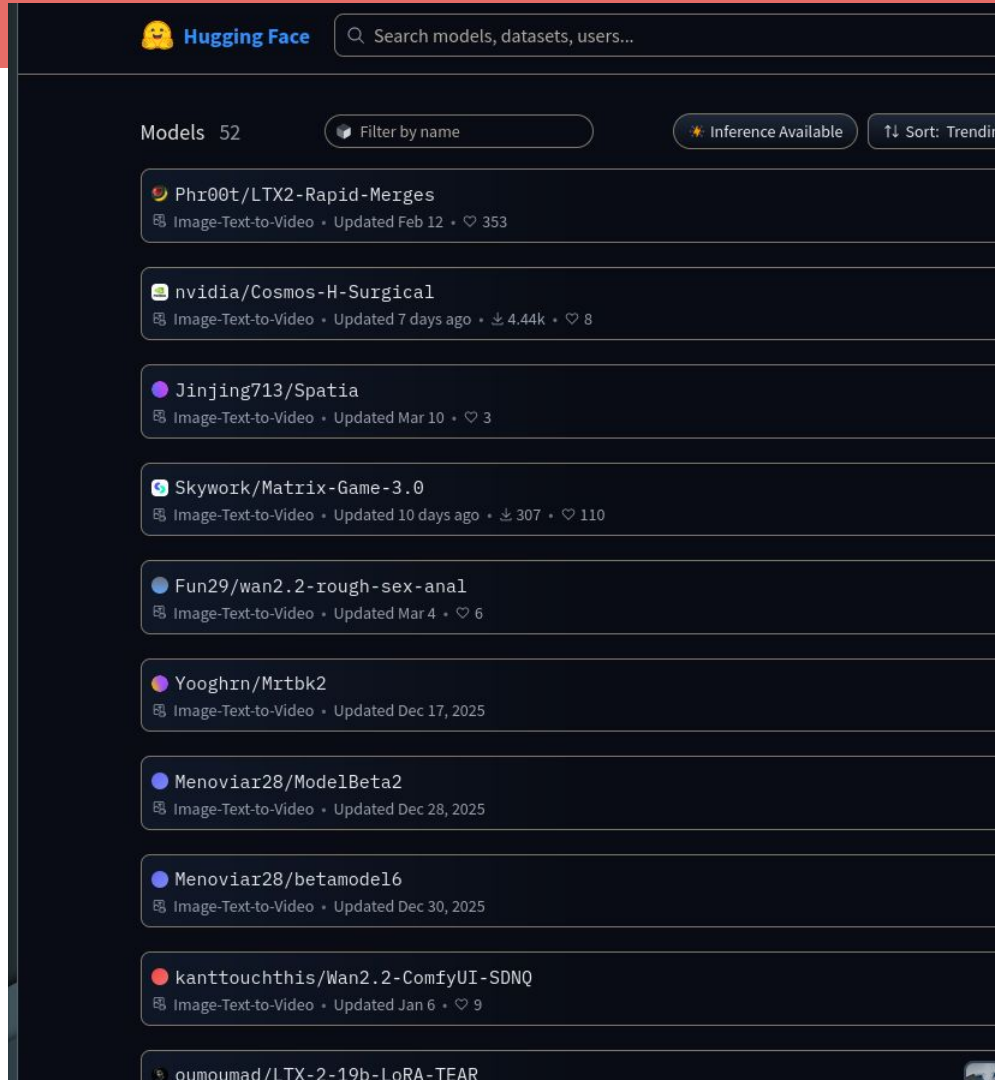
How can I use AI to generate nudes on my comput...



# Model Access Providers

Two-thirds of the top trending image-and-text-to-video models on Hugging Face were devoted to generating sexual scenes, their names describing specific acts.

Tattle



# What Can Anyone Do About This?

Tattle

# The Most Obvious/ Common Response

Takedowns

- Technical and legal precedent for it.
- Known issues: enforcement.
- Post harm intervention.

Tattle

Core infrastructure (cloudflare, DNS) providers enable all these layers

### Discovery Channels

Platforms through which people can find the services that nudify or generate sexual scenes.

Google Search · Play Store · Social Media · ChatGPT / LLMs · YouTube · ...

### Graphical User Interface

The easy to use modality via which non technical users can upload images to generate synthetic content

Telegram Bot · Phone App · Websites · Discord Bot · X/Grok · ...

### Model Access Providers

Interfaces through which more technical users can access models and build on top of AI models. Providers might also support model downloads, gamification and monetisation features.

Hugging Face · GitHub · CivitAI · REST API · ...

### Machine Learning Models

Models are trained on individual/corporate hardware and datasets. Large base media models are fine tuned on specific sexual scenes.

Finetune (LoRA) · Stable Diffusion · Flux · Wan2.2 · Grok · ...

### Datasets

Images and videos are collected and annotated. Datasets enable base model training and fine-tuning

Personal data from Social Media · Porn Videos · LAION-5B · YouTube videos · ...

# Datasets

- Remove nudity from datasets.
  - This results in overall performance decline.
- Poison the data:
  - End users: so that the image can't be manipulated
  - Adult content creators: so that their content can't be used in training

Tattle

# Model Developers

Safety Evaluations  Guardrails

But with open weight models these guardrails can be removed.

Building open weight models with tamper-proof guardrails is an unsolved problem.

Tattle

# Model Developers

- Release under a Responsible AI License that prohibits NCSI use.
  - hard to enforce
  - violates free software principles



Tattle

# Model Access Providers

- Remove sexual content generating models
- Demonetize sexual content generating models
- Ban use of real human images (Civitai)

Tattle

# Takeaway

**Open Weight models are a (Big) contributor to NCSI**

**With open weight models: Better mitigation upstream or downstream.**

**Open source advocates concerned about minimizing harm need to be invested in interventions elsewhere in the stack.**

Tattle

# Thank you!

# Questions?



Tattle