# Platforms or Individuals: Who should be held responsible for generating AI mature contents that may be illegal?

## Group A

This topic covers a range of mature content. Please be advised.

Trigger Warning: sexually explicit content, racial slurs

# Topic Overview

# What is AI-generated media?

AI-generated media, or synthetic media, is an umbrella term encompassing all media (text, image, video, voice, etc.) that is fully or partially generated by AI.

- characterized by a high degree of of realism, often making it indistinguishable from real media

- **Deepfake** (deep learning + fake)
  - uses a facial recognition algorithm and a deep learning neural network to swap someone in an image or video with another person's likeness
  - evolved from facial reanimation technology intended for applications in movie dubbing
- **AI-generated text** (e.g. ChatGPT, Jasper, Frase)
  - requires minimal human input; produces high-quality text, can be made to imitate a specific writing style
- **Synthetic video** (e.g. Synthesia, Synthesys, Rephrase.ai)
  - most commonly made with a life-like digital avatar and a text-to-speech engine, but can also bring still images, such as portraits, to life
  - reduces the time and cost that would be required to physically film a video

# What constitutes mature content?

**Mature content**: refers to material that is intended for audiences who are of a certain age or maturity level.

*The classification of content as "mature" can vary depending on the context and cultural norms. However, some common elements that are often associated with mature content include:*

# The question

When referring to mature content generated by AI, it typically means content produced by an artificial intelligence system and containing elements that may be considered adult content.

**Should AI platforms be held responsible for the mature content they generate, or should individuals be the ones held accountable?**

*Team B: Platforms should be held accountable.*

*Team C: Individuals should be held accountable.*

# Case Studies: People abusing platforms

# AI-generated hate speech language model

Yannic Kilcher, an AI researcher and YouTuber, trained an AI on 3.3 million postings from 4chan's infamously Politically Incorrect /pol/ board.

Kilcher released the AI on the board after implementing the model in 10 bots, which resulted in a wave of hatred. The bots produced 15,000 posts in 24 hours that frequently featured or engaged with racist information.

He described the experiment as a "prank," not research. It serves as a reminder that trained AI is only as good as the data it is fed.

# Predators Are Abusing Generative AI

Child predators are using generative AI for serious text and image-based violations, including:

Production of guides on how to locate and groom vulnerable minors

Generation of scripts to communicate with and groom minors

Writing poetry and short stories that describe children in a sexual context

Modification and sexual distortion of existing images of children

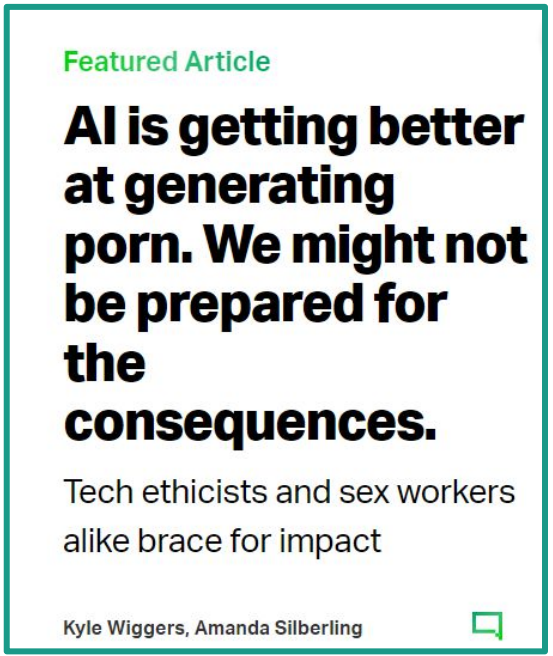
Creation of novel pseudo-photographic CSAM

While the challenge is great, effective moderation and risk mitigation are possible.

# Case Studies: Platforms

# Platforms emerging for AI porn

- AI platforms are emerging to explicitly create porn
- Platforms like “Porn Pen” allow users to create nude, AI-models for NSFW uses
- Porn Pen’s customizable models rival adult content creators online who create similar works on OnlyFans or ManyVids
- Porn Pen has been called by a PhD at University of Washington as “heteronormative”
- Sex workers fear harsh crackdowns on AI porn could lead to stricter legal hurdles that would make their work infeasible



# Unlawful Uses of Facial Recognition Tech

- The French Data Protection Authority fined Clearview AI 20 million euros for unlawful use
- Essentially, the AI was accessing sensitive data without consent for collection
- Clearview AI was found to be intrusive nature
- Even though the AI was programed to access public pictures on social media for facial recognition, it was also accessing private data

CNIL Fines Clearview AI 20 Million Euros for Unlawful Use of Facial Recognition Technology

Posted on October 24, 2022

POSTED IN ENFORCEMENT, EUROPEAN UNION, INTERNATIONAL, ONLINE PRIVACY

# Instagram's Algorithm Prefers A Little More Skin

- Third party researchers found that Instagram's algorithm prioritized showing images where users showed skin
- The researchers speculate that this algorithmic preference might stem from a wide usage of Instagram as "soft porn photos"
- While the algorithm learns from its users, it has in turned used that data to create its own biases

# Example: Getty Images sues AI art generator Stable Diffusion in the US for copyright infringement

Getty Images has filed a lawsuit in the US against Stability AI, creators of open-source AI art generator Stable Diffusion, escalating its legal battle against the firm.

The stock photography company is accusing Stability AI of “brazen infringement of Getty Images’ intellectual property on a staggering scale.” It claims that Stability AI copied more than 12 million images from its database “without permission ... or compensation ... as part of its efforts to build a competing business,” and that the startup has infringed on both the company’s copyright and trademark protections.


# Yifei Liu  becomes the "face" of the porn market, AI face replacement technology is abused

In recent years, with the rapid development of artificial intelligence technology, an increasing number of female celebrities have become "representatives" in the adult film industry. Liu Yifei is one of the most affected individuals. It has been reported that Liu Yifei's facial features have become a "standard feature" for adult film production companies. These websites use AI face-swapping technology to embed Liu Yifei's facial features into adult films, creating an illusion of apparent authenticity. Such behavior not only seriously harms Liu Yifei's reputation, but also goes against ethical standards and legal regulations.

# Case Studies: Gray Areas

# Example: AI-assisted clothing removal (Deepfake, etc.)

AI-generated nude photo of Chinese social media influencer sparks outrage, raises concerns of cybercrime

In a recent incident, a photo of a female social media influencer falsely depicted as naked in Guangzhou's metro station was circulated on social media. The image was fabricated using an AI-powered software that generates nude photos with just one click. However, the original photo, of a well-known female social media influencer wearing shorts and a vest, was tampered with and altered to create the fake nude image.

The fashion blogger, who has remained anonymous, has vowed to take legal action against those who created the fake photo.

# Example: AI-generated deep fake pornography scandal

Recently, AI-generated deep fake pornography has been heavily circulating online. One of the most recent scandals included a twitch streamer who goes by the name “Atrioc” , who went viral for looking at AI- generated deep fakes of other popular female streamers.

The streamer has since then step down from content creation and vowed to help combat the spread of such content by covering the legal cost of removing them from the internet.

The female users affected by the incident have spoken out about taking legal action against both the streamer as well as the hosts of the AI- generating website.

# Effects of such content on the victims

In the case presented, the affected female streamers spoke about some of the physiological damage the scandal has caused them. Some of  which include:

- Violation of their human rights
- Invasion of privacy
- Loss of self preservation
- Triggering of underlying body images and personal struggles

# Response of platform to the scandal

The platform **“Twitch”** that hosts the streamers talked about in the previous scandal put out a statement a month after the unfolding of the situation. Their statement addressed their stance against what they referred to as “synthetic non-consensual exploitative images” or NCEI in general. The main takeaways of the statement were:

- Consultation with experts in the field
- Two major updates in the platforms policy
- Announcement of a creator camp on March 14 to “help protect women streamers”

> 1. We’re updating our Adult Sexual Violence and Exploitation policy to make it more clear that intentionally promoting, creating, or sharing synthetic NCEI can result in an indefinite suspension on the first offense.
>
> 2. We’re updating our Adult Nudity policy to include synthetic NCEI. Even if that NCEI is shown only briefly, or, for example, shown to express your outrage or disapproval of the content, it will be removed and will result in an enforcement.

# Who's at fault? Users?

*Faulting the software and platform does not tackle the actual problem….*

Agency and Responsibility: AI is just a program/tool controlled by human

Lack of Consciousness: AI or the platform itself  does not have capability, or understanding to make moral/ethical decisions

Worries of legal problems as a result of user misuse might limit innovation  and discourage  development

# Who's at fault? The platform?

| | |
|---|---|
| 1 **Hosting and Distribution** | - When **no action to restrict distribution or remove content**, the platform can be held liable |
| 2 **Terms of Service Violation** | - When **failing to enforce its own policies**, possible negligence |
| 3 **Lack of Content Moderation** | - Platforms **expected to implement content moderation measures** to ensure the safety of users |
| 4 **Failure to Respond to Reports** | - When **failing to respond appropriately** to reports by e.g. removing content and punishing accordingly |

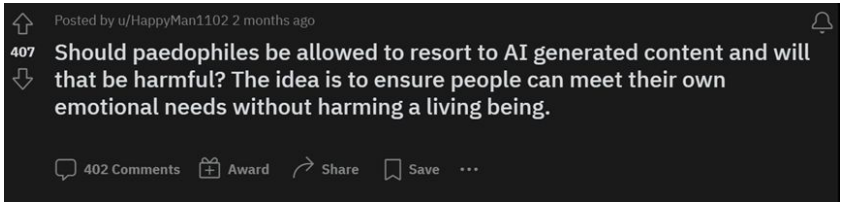
# Hot Debates on AI-generated mature content

For:

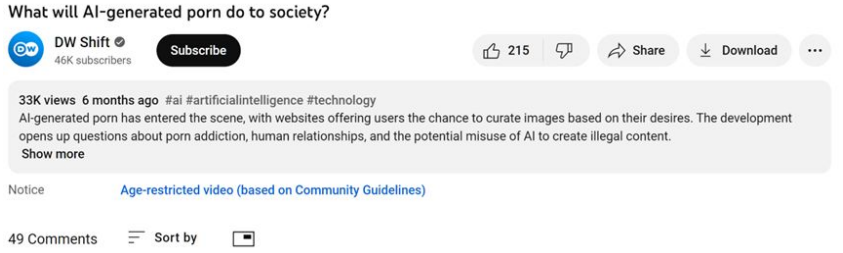
- Humans (especially women) will no longer be sexualized and reduced to how they look and please others.
- It can serve as a cop-out mechanism instead, while trying to get rid of the impulses
- It must be possible to do it without using actual child abuse imagery as a reference.
- Everything should be legal by default, only those things which probably harm others should be outlawed.
- It leads to a tax-friendly business.
- …

Against:

- AI uses information using the things it finds which it would and has already pulled from actual CP which does harm the child/children having more of their exploitation out there
- Regular porn has been shown not to be connected to increased probability of rape.
- Porn addiction has been shown to lead to escalation, escalation in what it takes to get off, and to more and more risky behaviour irl.
- AI porn will only advance the porn industry.
- …




https://www.reddit.com/r/morbidquestions/comments/123nba5/should_paedophiles_be_allowed_to_resort_to_ai/




https://www.youtube.com/watch?v=vvfyHlnl9tk&list=PLLhNQgjyovDI7znkG2FiqLWYpk0NpxfDi&index=1

# Debate Time!

# Argument flow:

- Topic overview → introduce types of platforms and intended use (deepfake for example)**Emily** and definitions of mature content/AI **Kinani** introduce the question and clarify meaning **Kinani**
- Case studies: people at fault → **Holyfield**'s Yannic case, **Holyfield**'s predator case
- Case studies: platform is at fault → **Audrey** will find a case study, **Rongqian**
- Cases where there is nuance (either or) → **Eman**'s case, **Holyfield**'s AI-assisted clothing removal case
- People at fault (accountability of people, malicious intent)→ **Denise**
- Platforms at fault (legal) → **Pablo**
- *Debate: population consensus*
- **Debate facilitators?**

Should AI platforms be held responsible for the mature content they generate, or should individuals be the ones held accountable?

Team B: Platforms should be held accountable. ; Team C: Individuals should be held accountable.