

Safe(r) Digital Intimacy Session I

Defense-in-Depth Against Image Based Sexual Abuse

Elissa M. Redmiles

elissa.redmiles@georgetown.edu

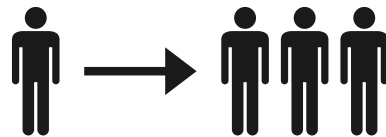
Trigger Warning: Sexual Violence



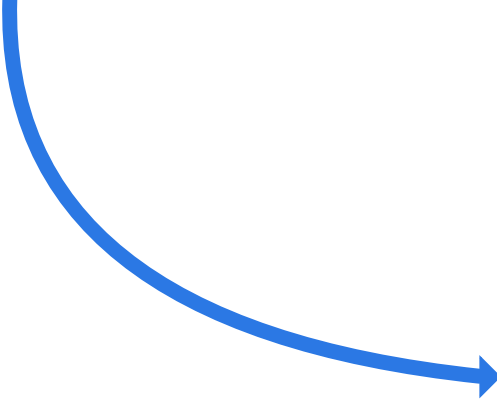
GEORGETOWN UNIVERSITY

[safe digital intimacy.org](https://safe.digitalintimacy.org)

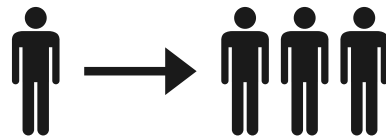
A friend uses AI to create
a synthetic video of you
dancing and sends it a
group chat with you in it



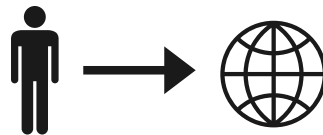
Do you find this acceptable?



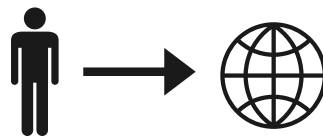
A friend uses AI to create a synthetic video of you dancing and sends it a group chat with you in it



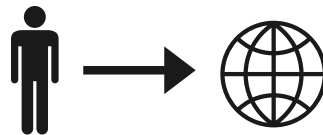
A friend uses AI to create
a synthetic video of you
dancing and **posts it
publicly**



A **stranger** uses AI to
create a synthetic video of
you dancing and posts it
publicly



A stranger uses AI to
create a synthetic video of
you **performing a sexual
act** and posts it publicly

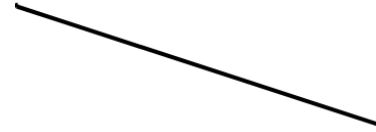


AI-generation of **synthetic non-consensual intimate imagery (SNCII)** is a **form of image based sexual abuse**.

I. CONTEXT SETTING

IBSA

Image-Based Sexual Abuse



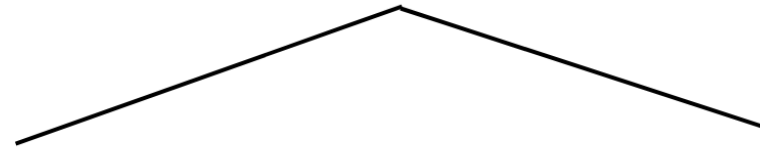
NCII

Non-consensual
Creation of
Intimate Imagery

I. CONTEXT SETTING

IBSA

Image-Based Sexual Abuse



NDII

Non-consensual
Distribution of
Intimate Imagery

NCII

Non-consensual
Creation of
Intimate Imagery

I. CONTEXT SETTING

***Intimate images** include images or videos that show a nude or semi-nude subject, contain intimate body parts, and/or intend to arouse.

I. CONTEXT SETTING

Image based sexual abuse
is a **violation in and of itself**

and can lead to....

I. CONTEXT SETTING

1

Mental health consequences

Suicide and other serious mental health consequences similar to other forms of sexual abuse (PTSD, depression)

2

Doxxing & online harassment

Increased risk of doxxing and other online hate and harassment. The risks are magnified for LGBTQ, women in restrictive countries and sex workers

3

(US) Legal Implications

Forty-eight US states have laws against non-consensual distribution of intimate imagery (NDII)

US federal law just passed against SNCL + 16 states have laws on SNCL.

Source: cybercivilrights.org/deep-fake-laws/

I. CONTEXT SETTING

Since 2018, I've been studying security for intimate content & interactions

in collaboration with

Hanna Barakat
Sarah A. Bargal
Catherine Barwulor
Natalie Grace Brigham
Kevin Butler
Ana-Maria Cretu
Anna Crowder
Adam Doupe
Diana Freed
Vaughn Hamilton
Eszter Hargittai
Tadayoshi Kohno

Cassidy Gibson
Rupayan Mallick
Michelle Mazurek
Allison McDonald
Daniel Olszewski
Lucy Qin
Florian Schaub
Ananta Soneji
Patrick Traynor
Carmela Troncoso
Sharon Wang
Miranda Wei



RollingStone
SCIENTIFIC
AMERICAN
BuzzFeed News
The Daily Dot

THE WALL STREET JOURNAL.
EL PAÍS
INDEPENDENT
WINK
NEWS
NETZPOLITIK.ORG

I. CONTEXT SETTING

Since 2018, I've been studying security for intimate content & interactions



Trauma Aware

All research team members are **trained in trauma aware research practice**

Mental health professional available for regular debriefs throughout research



Research Justice

Members of impacted communities are included as **peer researchers, ethics consultants & transcribers**

Results are reported back to community in accessible formats (one pagers, white papers, policy remarks, etc.)



Ethics

Labor (participants, researchers, etc.) is compensated

All studies are approved by IRB or local equivalent

agenda

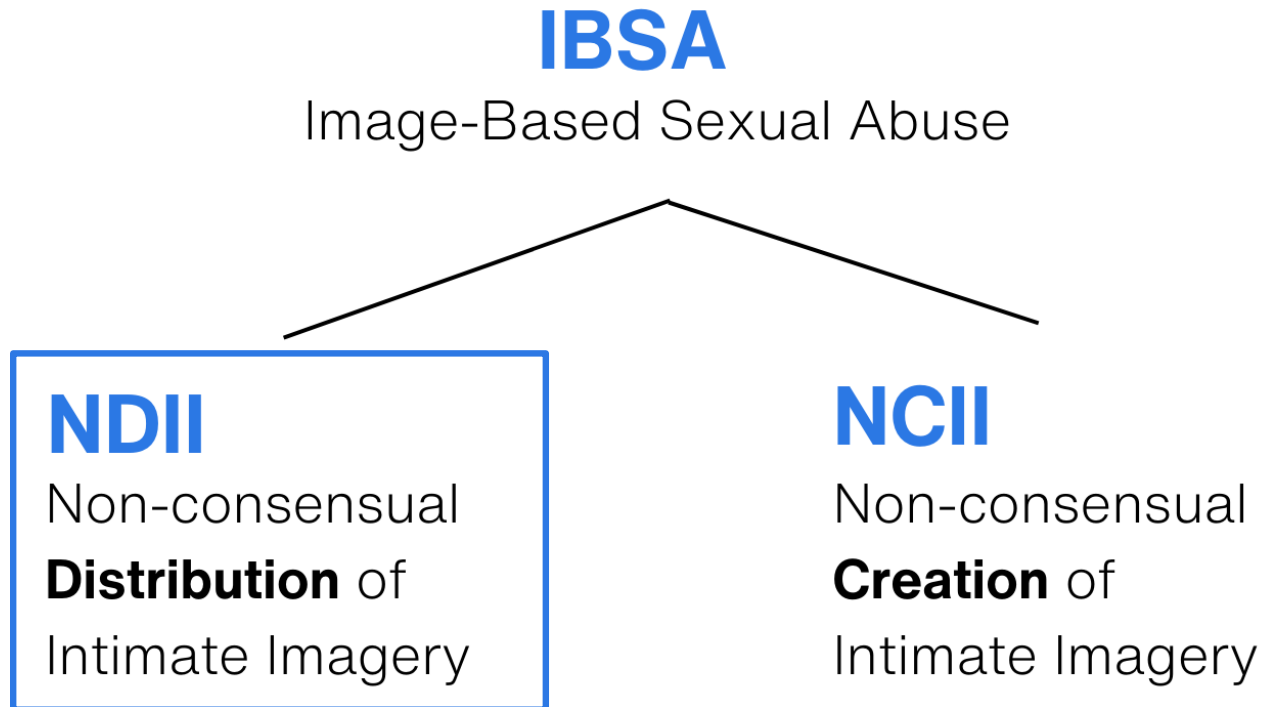
I. Context Setting

II. NDII: Intimate Sharing

III. SNCII: AI Generation

IV. Defining AI Safety: Mitigations & Gaps

II. NDII: Intimate Sharing



II. NDII: Intimate Sharing

“Did They Consent to That?”

Safer Digital Intimacy via Proactive
Protection Against Image-Based
Sexual Abuse

Lucy Qin, Vaughn, Sharon Wang,
Yigit Aydinalp, Marin Scarlett,
Elissa M. Redmiles

USENIX Security
2024

Research question

RQ

What are the contexts, technologies, threat models, and proactive defensive strategies involved in intimate content sharing?

II. NDII: Intimate Sharing

Qualitative Interviews

In 2023 with 52 adults in the EU who
share intimate content consensually

28 shared recreationally,
24 shared for commercial purposes
22 were victim-survivors of NDII



II. NDII: Intimate Sharing

People share intimate content with...



Strangers

(e.g., body positivity group
on social media)



New relationships

(e.g., someone
on a dating app)



Established

relationships
(e.g., dating, marriage)



Commercial Platforms

(OnlyFans, etc.)

II. NDII: Intimate Sharing

People share intimate content with...



Strangers

(e.g., body positivity group on social media)



New relationships

(e.g., someone on a dating app)



Established

relationships
(e.g., dating, marriage)



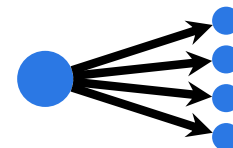
Commercial Platforms

(OnlyFans, etc.)

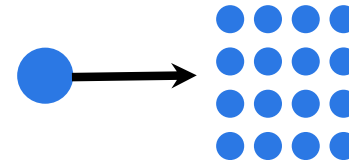
In different structures



1 to 1



1 to many



1 to a group

II. NDII: Intimate Sharing

“Early [in] my transition, I had the need to just share some pictures [to] feel more sexy. And so I created an account [on social content platform], that [account] is networked with mostly other trans women that also have these accounts for the exact same purpose.”

(Research Participant)

II. NDII: Intimate Sharing

“I’ve shared them with a lot of people. I’ve shared them with romantic partners, I’ve sent them to strangers on the internet...I have like a craving for praise. And so that motivates me to share, to share my body with people who might enjoy it...I just look for people [on social media platform] who are interested in seeing those kind of pictures.”

(Research Participant)

II. NDII: Intimate Sharing

Participants used 40+ platforms

Any platform that can create, share or store visual content **is likely used for intimate content**



Messaging
(apps and SMS)



**Social Media
Platforms**



**Hookup/
Dating
Apps**



**Adult Content
Platforms**



**File Share
Platforms**



Email

II. NDII: Intimate Sharing

“...I’ve got like Telegram, I’ve got Signal, I’ve got WhatsApp, I’ve got Kik, you name it, I’ve got it... occasionally, if there’s something really large, that requires drop boxes and that sort of thing. And I’ll do that”

(Research Participant)

II. NDII: Intimate Sharing

Participants used 40+ platforms

Any platform that can create, share or store visual content **is likely used for intimate content**

More on security / tech use in sex work tomorrow or later this session



Messaging
(apps and SMS)



Social Media Platforms



**Hookup/
Dating Apps**



Adult Content Platforms



File Share Platforms



Email

II. NDII: Intimate Sharing

What are
the risks of sharing
intimate content?

Recipient

Non-consensual distribution of intimate imagery

II. NDII: Intimate Sharing

What are the risks of sharing intimate content?

Recipient

Non-consensual distribution of intimate imagery

Non-Recipient

Device Sharing: someone sharing device accidentally finds content

Shoulder Surfing: someone looking over a recipient's shoulder in public accidentally sees content

Hacking: breach of the platform databases storing content

Insider Threat: company employee viewing content illicitly

II. NDII: Intimate Sharing

To protect themselves, participants used **technological strategies** when available and **interpersonal strategies** to fill in the gaps

II. NDII: Intimate Sharing

Defending against recipient resharing

Before Sharing



Rule
Setting



Screening/
Vetting

II. NDII: Intimate Sharing

Defending against recipient resharing

Before Sharing



Rule
Setting



Screening/
Vetting

While Sharing



Expiring
Messages



Screenshot
Notifications



Watermarks

II. NDII: Intimate Sharing

Defending against recipient resharing

Before Sharing



Rule
Setting



Screening/
Vetting

While Sharing



Expiring
Messages



Screenshot
Notifications



Watermarks

After Sharing



Deletion
Request



Message
Unsend

II. NDII: Intimate Sharing

Defending against recipient resharing

Before Sharing



Rule
Setting



Screening/
Vetting

While Sharing



Expiring
Messages



Screenshot
Notifications



Watermarks

After Sharing



Deletion
Request



Message
Unsend

Defending against identification



Removing
Identifying
Features



Metadata
Removal

Friction Matters

Deterrence is a key principle of crime prevention.
Adding friction can defend against average- or low-effort abusers.

“No matter how much features people put into safety, there’s always always always going to be a risk...[but] as long as the features and the way you’re doing it has the minimum level of safety, that will [stop] most people.”

(Research Participant)

II. NDII: Intimate Sharing

What are the
perceived risks
of sharing?

Recipient

Non-consensual distribution of intimate imagery

II. NDII: Intimate Sharing

What are the perceived risks of sharing?

Recipient

Non-consensual distribution of intimate imagery

Non-Recipient

Device Sharing: someone sharing device accidentally finds content

Shoulder Surfing: someone looking over a recipient's shoulder in public accidentally sees content

Hacking: breach of the platform databases storing content

Insider Threat: company employee viewing content illicitly

II. NDII: Intimate Sharing

To protect themselves, participants used **technological strategies** when available and **interpersonal strategies** to fill in the gaps

II. NDII: Intimate Sharing

Defending against recipient resharing

Before Sharing



Rule
Setting



Screening/
Vetting

While Sharing



Expiring
Messages



Screenshot
Notifications



Watermarks

After Sharing



Deletion
Request



Message
Unsend

II. NDII: Intimate Sharing

Defending against recipient resharing

Before Sharing



Rule
Setting



Screening/
Vetting

While Sharing



Expiring
Messages



Screenshot
Notifications



Watermarks

After Sharing



Deletion
Request



Message
Unsend

Defending against identification



Removing
Identifying
Features



Metadata
Removal

Friction Matters

Deterrence is a key principle of crime prevention.
Adding friction can defend against average- or low-effort abusers.

“No matter how much features people put into safety, there’s always always always going to be a risk...[but] as long as the features and the way you’re doing it has the minimum level of safety, that will [stop] most people.”

(Research Participant)

safe digital intimacy.org

	 Google Messages	 Apple iMessage	 Discord	 Grindr	 Instagram DM	 Signal	 Snapchat	 WhatsApp
Screenshot Prevention	Not available	Not available	Not available	Variant of desired feature available	Not available	Not available	Not available	Variant of desired feature available
Download Prevention	Not available	Not available	Not available	Variant of desired feature available	Variant of desired feature available	Variant of desired feature available	Variant of desired feature available	Variant of desired feature available
Screenshot Notification	Not available	Not available	Not available	Not available	Variant of desired feature available	Not available	Variant of desired feature available	Not available
Watermarking	Paid or partial implementation	Not available	Not available	Not available	Not available	Paid or partial implementation	Paid or partial implementation	Paid or partial implementation
Mask on Send	Not available	Paid or partial implementation	Variant of desired feature available	Variant of desired feature available	Not available	Variant of desired feature available	Variant of desired feature available	Variant of desired feature available
Automatic Metadata Removal	Variant of desired feature available	Not available	Variant of desired feature available	Variant of desired feature available	Variant of desired feature available	Variant of desired feature available	Variant of desired feature available	Variant of desired feature available
Blur Identifying Features	Paid or partial implementation	Not available	Not available	Not available	Not available	Variant of desired feature available	Paid or partial implementation	Paid or partial implementation
Message Unsend*	Not available	Variant of desired feature available	Variant of desired feature available	Paid or partial implementation	Variant of desired feature available	Variant of desired feature available	Variant of desired feature available	Variant of desired feature available
Expiring Messages*	Not available	Not available	Paid or partial implementation	Paid or partial implementation	Variant of desired feature available	Variant of desired feature available	Variant of desired feature available	Variant of desired feature available

Not available

Variant of desired feature available

Paid or partial implementation

VI. Sociotechnical Mitigations & Gaps

What product changes would make for safe(r) digital intimacy?

Increase availability & visibility of existing features

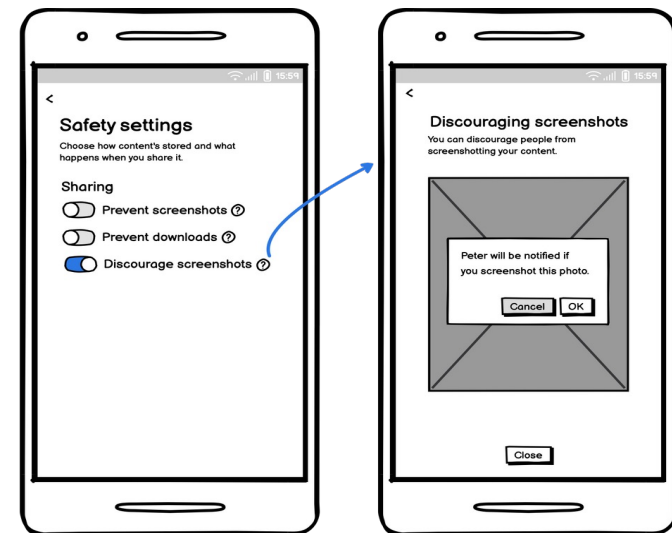
	Google Messages	Apple iMessage	Discord	Grindr	Instagram DM	Signal	Snapchat	WhatsApp
Screenshot Prevention								
Download Prevention								
Screenshot Notification								
Watermarking								
Mask on Send								
Automatic Metadata Removal								
Blur Identifying Features								
Message Unsend*								
Expiring Messages*								

Not available

Variant of desired feature available

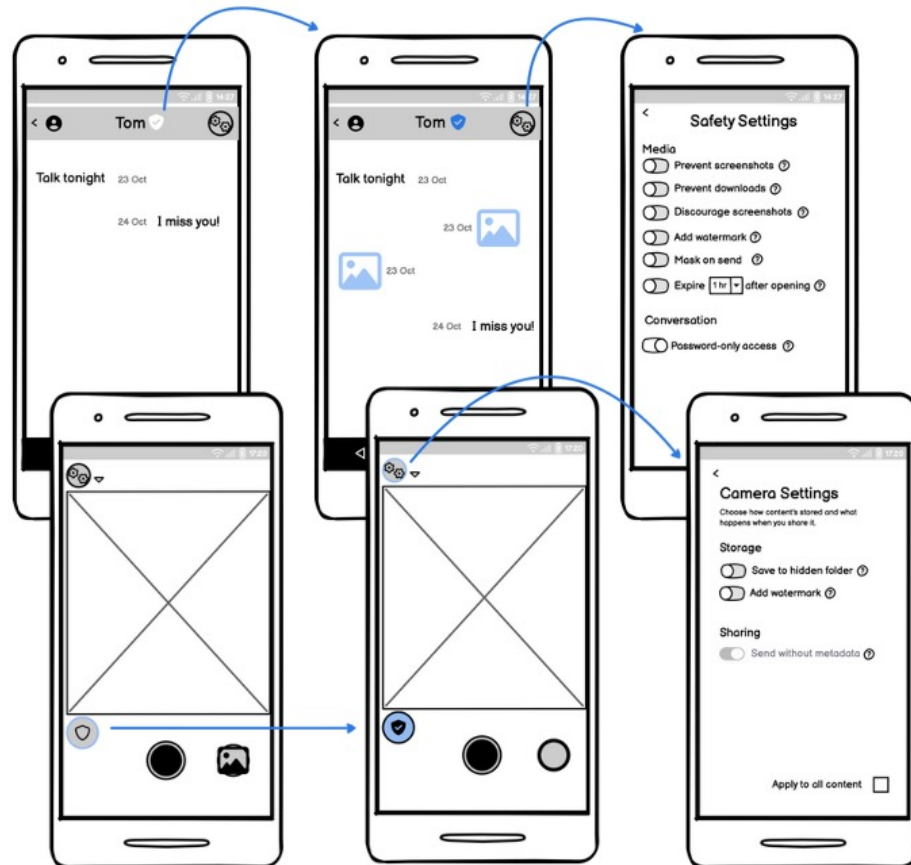
Paid or partial implementation

Content-level data control & presets



Our Participant's Vision:

one-click “intimate picture mode”
or a control panel of settings with
save-able presets for different
levels of trust (e.g., established
relationships vs. strangers)



IV. SOCIOTECHNICAL MITIGATIONS & GAPS

Open Question:

How can we **balance content control
with harm documentation?**

II. INTIMATE SHARING PERSONAS

Peter



- Peter shares intimate content with potential partners, including strangers, which is a norm on the dating apps he uses for meeting other gay men.
- Since he often shares intimate content with strangers, he's worried about one of them re-sharing his images publicly (such as on an online forum) without his consent.

IV. SOCIOTECHNICAL MITIGATIONS & GAPS

Open Question: How can we **balance content control with harm documentation?**

New design ideas should be explored to resolve the desire for features such as ephemerality while allowing senders to maintain documentation in case of harm

"[Expiring messages] sort of invite people trying to screenshot... it almost creates this environment of, oh, this is secret, which sort of invites people to be like, Oh, I'm going to try and hold on to it"
-P31

Ongoing Qualitative Interviews

**In 2023 & 2025 with 18 organizations
including government offices from
15 countries that take down content**

Africa (1)
Asia (2)
Europe (11)
North America (1)
Oceania (2)
South America (2)

Work led by Lucy Qin (Postdoctoral Researcher) in collaboration with Diana Freed (Brown) & Sharon Wang (UW Psyc)

II. NDII: Intimate Sharing

People share intimate content with...



Strangers

(e.g., body positivity group on social media)



New relationships

(e.g., someone on a dating app)



Established

relationships
(e.g., dating, marriage)



Commercial Platforms

(OnlyFans, etc.)

and as a result of coordinated attacks

II. NDII: Intimate Sharing

People share intimate content with...



Strangers

(e.g., body positivity group on social media)



New relationships

(e.g., someone on a dating app)



Established

relationships
(e.g., dating, marriage)



Commercial Platforms

(OnlyFans, etc.)

and as a result of coordinated attacks



Extortion

victim targeted via social media or dating apps

“almost like a call centre, in these places where they are just reaching out to hundreds and hundreds of men”

“they target like a lot of women at the same time. So it’s not... it’s not only the ex-partner...but in many cases, it’s more like coordinated”

(Organizational Research participants)

II. NDII: Intimate Sharing

People share intimate content with...



Strangers

(e.g., body positivity group on social media)



New relationships

(e.g., someone on a dating app)



Established

relationships
(e.g., dating, marriage)



Commercial Platforms

(OnlyFans, etc.)

and as a result of coordinated attacks



Extortion

victim targeted via social media or dating apps

1. “**They think they’re getting in contact with a young woman that is interested in them**, maybe on a dating app or on Instagram.
2. And then they have a flirt, maybe during a week or so, write each other back and forth.
3. It develops into sharing some intimate images or being naked in a sexual situation on video cam together.
4. And then it switches. The blackmail starts”

II. NDII: Intimate Sharing

People share intimate content with...



Strangers

(e.g., body positivity group on social media)



New relationships

(e.g., someone on a dating app)



Established

relationships
(e.g., dating, marriage)



Commercial Platforms

(OnlyFans, etc.)

and as a result of coordinated attacks



Extortion

victim targeted via social media or dating apps

1. “They think they’re getting in contact with a young woman that is interested in them, maybe on a dating app or on Instagram.
2. And **then they have a flirt**, maybe during a week or so, write each other back and forth.
3. It develops into sharing some intimate images or being naked in a sexual situation on video cam together.
4. And then it switches. The blackmail starts”

II. NDII: Intimate Sharing

People share intimate content with...



Strangers

(e.g., body positivity group on social media)



New relationships

(e.g., someone on a dating app)



Established

relationships
(e.g., dating, marriage)



Commercial Platforms

(OnlyFans, etc.)

and as a result of coordinated attacks



Extortion

victim targeted via social media or dating apps

1. “They think they’re getting in contact with a young woman that is interested in them, maybe on a dating app or on Instagram.
2. And then they have a flirt, maybe during a week or so, write each other back and forth.
3. **it develops into sharing some intimate images** or being naked in a sexual situation on video cam together.
4. And then it switches. The blackmail starts”

II. NDII: Intimate Sharing

People share intimate content with...



Strangers

(e.g., body positivity group on social media)



New relationships

(e.g., someone on a dating app)



Established

relationships
(e.g., dating, marriage)



Commercial Platforms

(OnlyFans, etc.)

and as a result of coordinated attacks



Extortion

victim targeted via social media or dating apps

1. “They think they’re getting in contact with a young woman that is interested in them, maybe on a dating app or on Instagram.
2. And then they have a flirt, maybe during a week or so, write each other back and forth.
3. it develops into sharing some intimate images or being naked in a sexual situation on video cam together.
4. **And then it switches. The blackmail starts”**

II. NDII: Intimate Sharing

People share intimate content with...



Strangers

(e.g., body positivity group on social media)



New relationships

(e.g., someone on a dating app)



Established

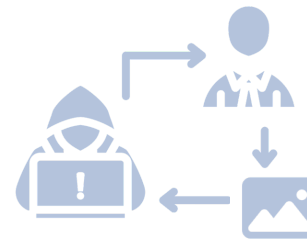
relationships
(e.g., dating, marriage)



Commercial Platforms

(OnlyFans, etc.)

and as a result of coordinated attacks



Extortion

victim targeted via social media or dating apps



Account Compromise

“So they steal the password, you know, enter to their account and find the images. So in many cases, young women don’t send these photos to anyone, only take photos for herself, let’s say, but these perpetrators steal them”

(Organizational Research participant)

Critical Need: Content Takedown

“The immediate overriding need is simply get the images down. It’s very rare that someone in crisis is asking for a lawyer, or police or even a therapist or anything like that, it is simply get the images down”

(Organizational Research participant)

II. NDII: Intimate Sharing

The process of
NCII takedown from
online platforms is
time consuming,
labor intensive,
and emotionally taxing.

II. NDII: Intimate Sharing

“[the process] is like pure agony”

(Research participant)

“Our practitioners spend four to six hours every day, manually searching and removing content, and checking that the content’s been removed. It takes so long . . . It’s mentally so draining.”

(Research participant: Organization that supports NCII victim-survivors)

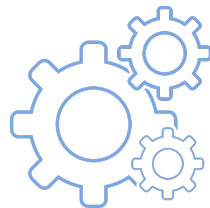
Building Tech to Evaluate IBSA takedown performance across platforms

Our goal is to...

empower victim-survivors and support organizations to track nonconsensual intimate content takedown from platforms.

advocate for platforms to offer this transparency directly.

Our team is building...



IBSA
takedown
tracker

The takedown tracker will...

- 1 **Automate manual processes** of checking if reported nonconsensual intimate content has been removed.
- 2 **Collect transparency data** on the efficacy and efficiency of takedowns.
- 3 **Highlight industry leaders & failings** in the mitigation of IBSA.



Georgetown
University

Lucy Qin
Elissa Redmiles



NEW YORK UNIVERSITY

Damon McCoy
Tobias Lauinger

MAX PLANCK
GESELLSCHAFT



Yixin Zou

Using DSA Article 40: Data Access and Scrutiny

1 To cross-validate takedown performance data

- We will use takedown tracker data to validate data made available through compliance with DSA Article 40 on takedowns of IBSA content from platforms.

2 To improve direct transparency about IBSA takedowns



Georgetown
University

Lucy Qin
Elissa Redmiles



NEW YORK UNIVERSITY

Damon McCoy
Tobias Lauinger

MAX PLANCK
GESELLSCHAFT



Yixin Zou

IV. Defining AI Safety: Mitigations & Gaps

AI Threat:

Ongoing **discovery & removal of NCII**

AI Threat: Ongoing **discovery & removal of NCII**

Only SoTA solution: Perceptual hashing
(aka PhotoDNA)

AI Threat: Ongoing **discovery & removal** of NCII

Only SoTA solution: Perceptual hashing
(aka PhotoDNA)

**“What if [a hash] could be
traced back to me?”**

(Research participant: Organization that supports NCII victim-survivors)

IV. Defining AI Safety: Mitigations & Gaps

AI Threat: Ongoing **discovery & removal** of NCII

Only SoTA solution: Perceptual hashing
(aka PhotoDNA)

Perceptual Hash Inversion Attacks on Image-Based Sexual Abuse Removal Tools

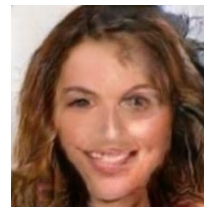
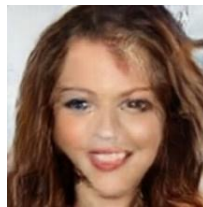
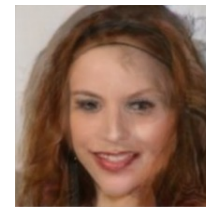
Sophie Hawkes  | Royal Holloway, University of London
Christian Weinert  | Royal Holloway, University of London
Teresa Almeida  | Umeå University and Interactive Technologies Institute/LARSyS
Maryam Mehrnezhad  | Royal Holloway, University of London

Source: Hawkes et al. IEEE S&P Magazine 2025

IV. Defining AI Safety: Mitigations & Gaps

Which of the following pictures, if any, would you feel comfortable with the social media platform storing so they can make sure the image is never posted again? **[select all that apply]**

Imagine that the following pictures were of you and imagine that instead of showing just your face, these pictures showed you performing a sexual act.

Original**Hawkes et al. reconstructions****Possible future result**

Defining AI Safety

Recognizability:

At what point is an image an image of you?

Recognizability to self vs. close contact vs. acquaintance vs. stranger

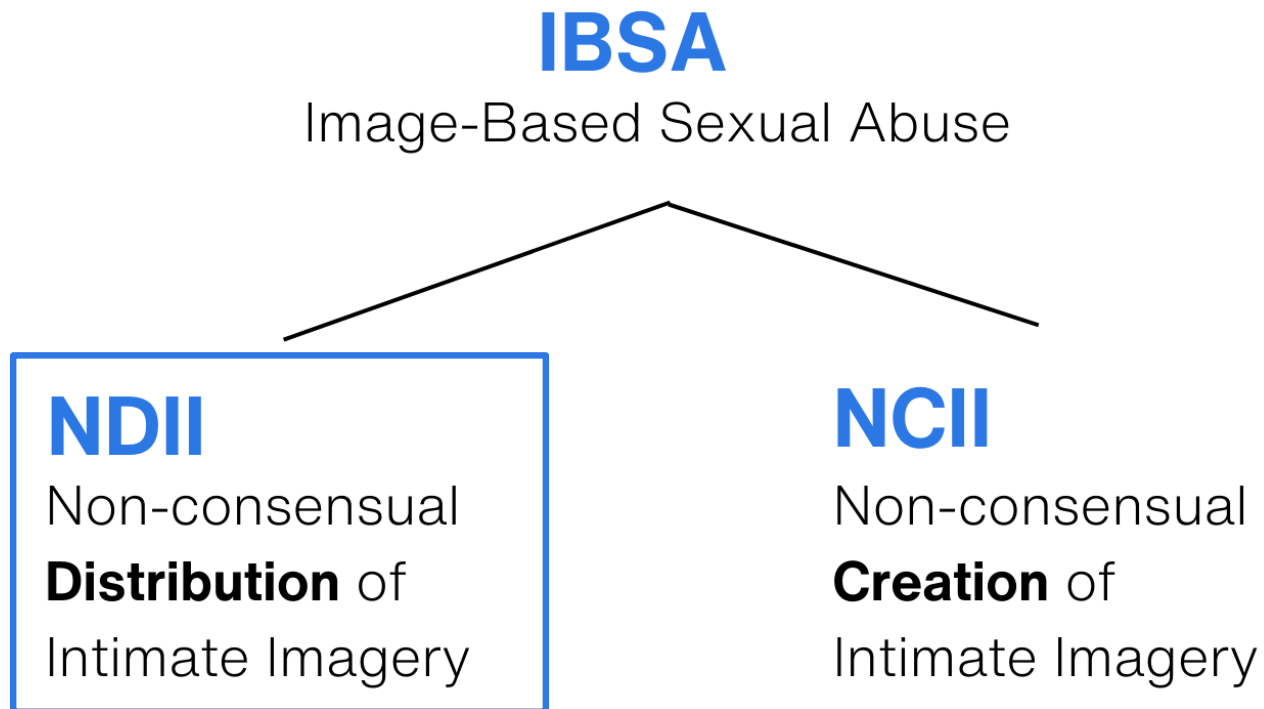
How does context change the threshold?

Defining AI Safety

Safety Theater vs. Threat Dramatics:

At what point is a tool doing more harm than good, even in the absence of other alternatives?

II. NDII: Intimate Sharing



IBSA

Image-Based Sexual Abuse

NDII

Non-consensual
Distribution of
Intimate Imagery

NCII

Non-consensual
Creation of
Intimate Imagery

III. SNCII: AI Generation

1

“Nudification” applications

These websites and applications purport to use computer vision techniques to undress the subject of an uploaded image

2

Generative Models

People may use generative models **hosted online** to generate SNCII of an individual by e.g., inputting a URL to an image of them as part of their prompt or they may run and fine tune their own **local models**

3

Hiring experts: Face swap + custom toolkits

Perpetrators may also hire experts who use a combination of face swapping, generative models, and image editing tools to create SNCII of a target.

III. SNCII: AI Generation

1

“Nudification” applications

These websites and applications purport to use computer vision techniques to undress the subject of an uploaded image

2

Generative Models

People may use generative models **hosted online** to generate SNCII of an individual by e.g., inputting a URL to an image of them as part of their prompt or they may run and fine tune their own **local models**

3

Hiring experts: Face swap + custom toolkits

Perpetrators may also hire experts who use a combination of face swapping, generative models, and image editing tools to create SNCII of a target.

III. SNCII: AI Generation

Analyzing the AI Nudification Application Ecosystem

Cassidy Gibson, Daniel Olszewski, Natalie Grace Brigham,
Anna Crowder, Kevin R.B. Butler, Patrick Traynor,
Elissa M. Redmiles, Tadayoshi Kohno

USENIX Security 2025

Research questions

- RQ1** How do nudification applications **position themselves to clients** via text and visual descriptions?
- RQ2** What **features** do nudification applications advertise?
- RQ3** How do nudification applications **monetize**?

APPLICATION WALKTHROUGH METHOD

20 nudification websites identified via NGOs & online lists

The core of this method involves the step-by-step observation and documentation of an app's screens, features and flows of activity – slowing down the mundane actions and interactions that form part of normal app use in order to make them salient and therefore available for critical analysis

III. SNCII: AI Generation

#1 AI Nude
Generator!

Undress your
girlfriend in one click

The MOST Realistic
Nudes

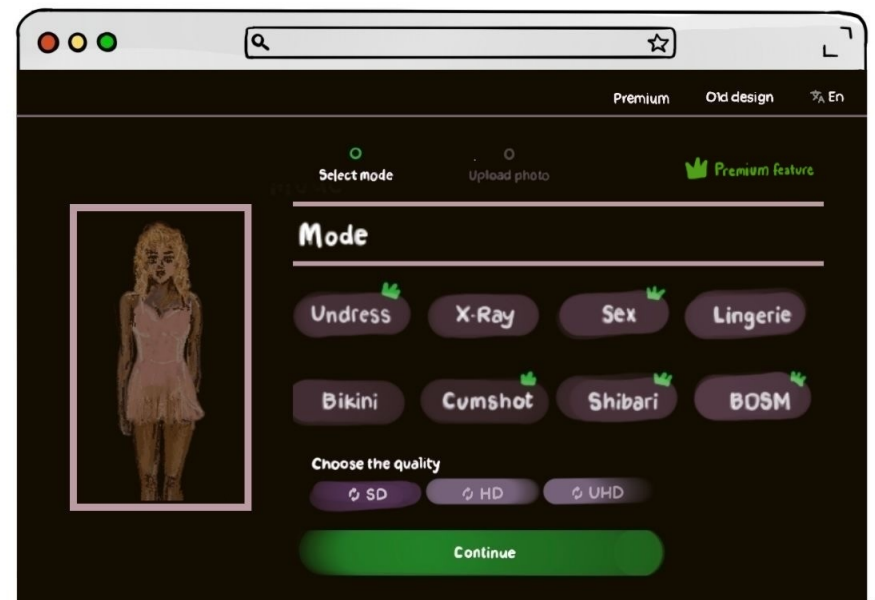
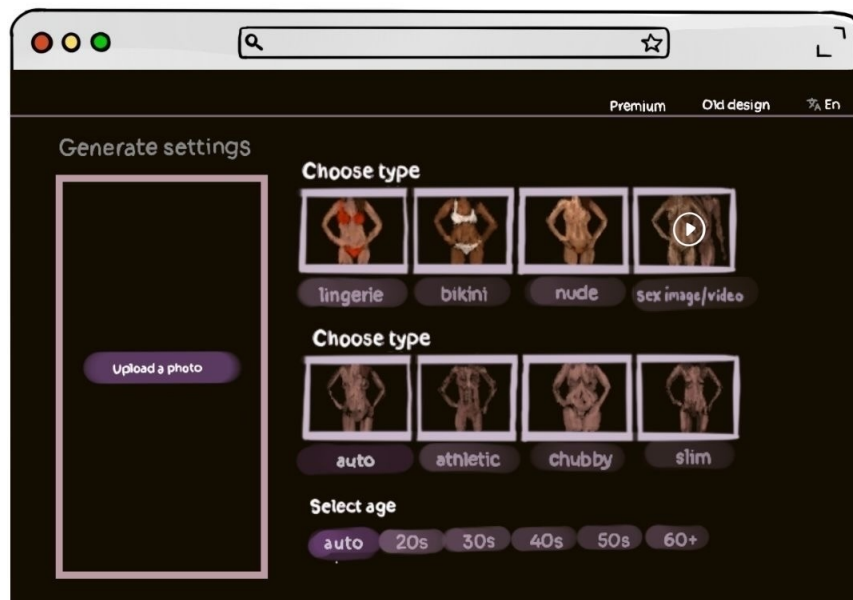
Make your own AI
Porn Video!

Our model has over 4370
Hours Trained on over
1,000,000 photos!

We don't save ANY
of your data

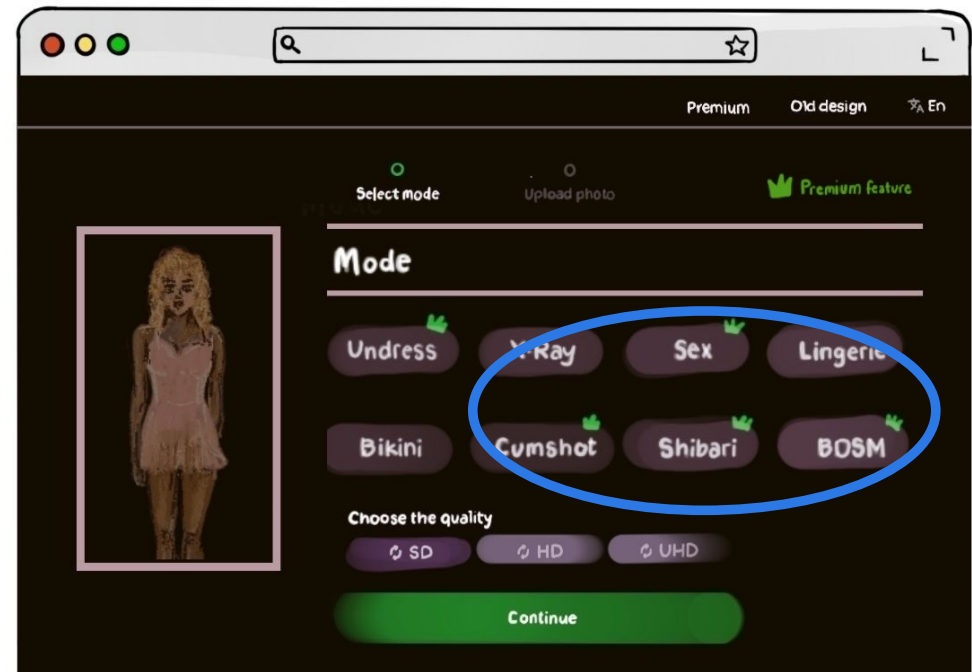
Undress ANYONE at
a click of a button

III. SNCII: AI Generation



Key findings

- 19 out of 20 applications **specialize in the non-consensual undressing of women**
- In addition to undressing, half of the applications purported to be able to **put the image subject in sexual scenarios**



Key findings

- 19 out of 20 applications specialize in the non-consensual undressing of women
- In addition to undressing, half of the applications purported to be able to put the image subject in sexual scenarios
- Many platforms **advertise through peer-to-peer methods:** affiliate programs, referral links, white labelling via APIs

Key findings

- 19 out of 20 applications specialize in the non-consensual undressing of women
- In addition to undressing, half of the applications purported to be able to put the image subject in sexual scenarios
- Many platforms advertise through peer-to-peer methods: affiliate programs, referral links, white labelling via APIs
- Average **price of ~\$0.34 per image**
[Han et al. find custom *video* deepfakes average \$87.50]

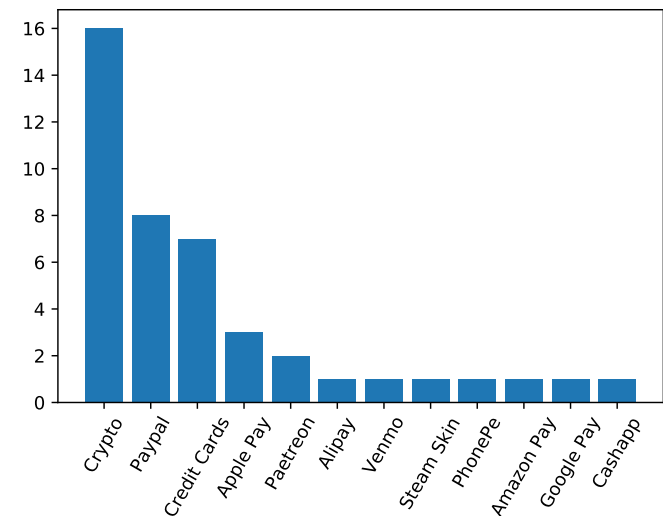
IV. Defining AI Safety: Mitigations & Gaps

Defense-in-Depth

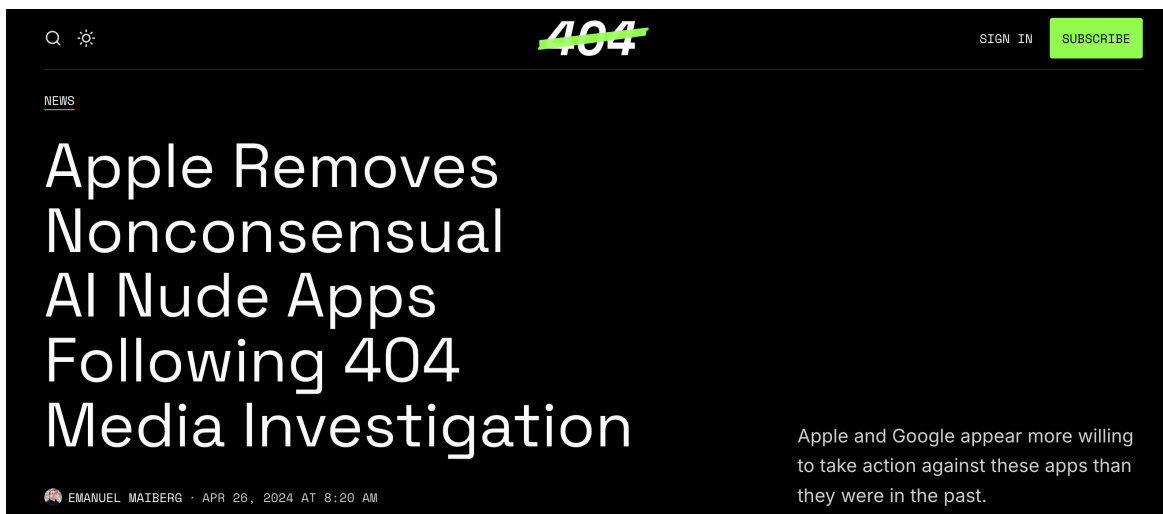
TARGET THE CREATION VALUE CHAIN

Many stakeholders:

- Payment processors
- Those whose software powers these sites
- Mainstream platforms where ads are run or search results are returned
- App stores & web hosts



IV. Defining AI Safety: Mitigations & Gaps



“The problem with apps is that they have this dual-use front where they present on the app store as a fun way to face swap, but then they are marketing on Meta as their primary purpose being nudification. So when these apps come up for review on the Apple or Google store, they don't necessarily have the wherewithal to ban them,”

—Alexios Mantzarlis

Defining AI Safety

Dual Use:

Can we create pipelines that identify misuse using contextual signals (ad language, search terms)?

IV. Defining AI Safety: Mitigations & Gaps

Defense-in-Depth **POISON INPUT IMAGES**



Shawn Shan, Ben Zhao et al. @ UChicago are exploring whether their approach for poisoning artists images can poison nudification systems

III. SNCII: AI Generation

1

“Nudification” applications

These websites and applications purport to use computer vision techniques to undress the subject of an uploaded image

2

Generative Models

People may use generative models **hosted online** to generate SNCII of an individual by e.g., inputting a URL to an image of them as part of their prompt or they may run and fine tune their own **local models**

III. SNCII: AI Generation

1

“Nudification” applications

These websites and applications purport to use computer vision techniques to undress the subject of an uploaded image

2

Generative Models

People may use generative models **hosted online** to generate SNCII of an individual by e.g., inputting a URL to an image of them as part of their prompt or they may run and fine tune their own **local models**

"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

IV. Defining AI Safety: Mitigations & Gaps

Defense-in-Depth **FORECASTING PERPETRATION**

Defense-in-Depth

FORECASTING PERPETRATION

Unintentionally-generated text-based SNCII creation

A user (not necessarily an adversary) prompts a text-to-image AI generator to generate an intimate image.

The AI generates an image of a person that matches the likeness of a real person, e.g., because of overtraining on images of that person.

Defense-in-Depth **RISK ASSESSMENT**

Approximation to upper bound risk of unintentional attack:

Image-extracted text-based SNCII creation

A UI bound attacker

- (1) uploads an image to a service that extracts a detailed text description of the target in the image and then
- (2) uses the resulting text to adversarially prompt a generative system for SNCII of the target.

Defense-in-Depth **RISK ASSESSMENT & FORECASTING PERPETRATION**

Approximation to upper bound risk of unintentional attack:-

Image-extracted text-based SNCII creation.

A UI bound attacker

- (1) uploads an image to a service that extracts a detailed text description of the target in the image and then
- (2) uses the resulting text to adversarially prompt a generative system for SNCII of the target.



“An opt-in service with a third-party vendor so individuals can register a privacy-preserving face hash and have future uploads blocked at submission.”

-- 404 Media quoting
CivitAI CEO Justin Maier

IV. Defining AI Safety: Mitigations & Gaps

Defense-in-Depth **RISK ASSESSMENT**

Approximation to upper bound risk of un

Image-extracted text-based SNCII creat
A UI bound attacker

- (1) uploads an image to a service that e
- text description of the target in the ima
- (2) uses the resulting text to adversarial
- generative system for SNCII of the targ

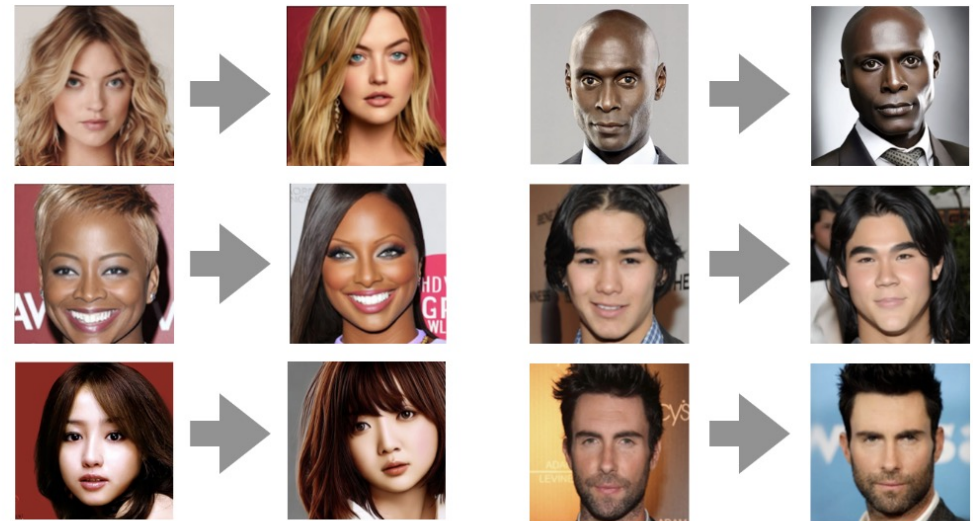


Figure 2: **Preliminary Work.** Generated images using encoded image embeddings.

IV. Defining AI Safety: Mitigations & Gaps

Red-Teaming Generative T2I diffusion models for propensity to generate digital replicas from English-language prompts

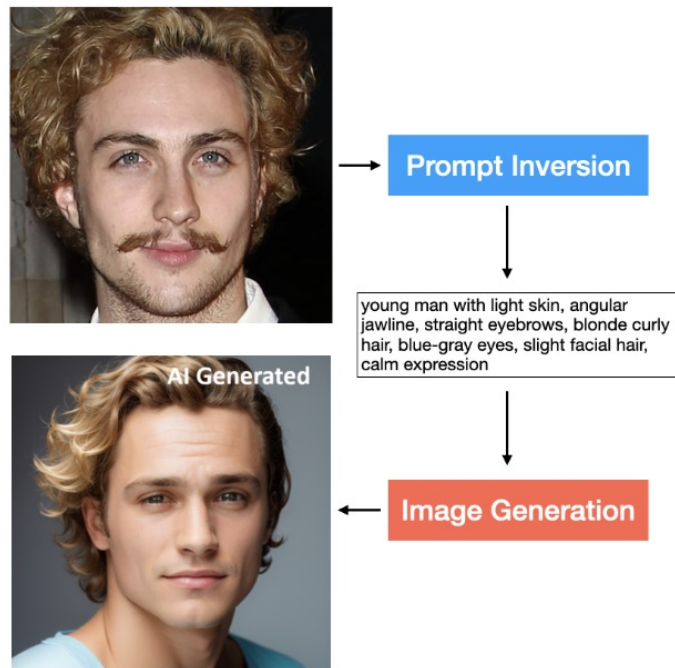
Rupayan Mallick, Amro Abdalla, Mahsa Khoshnoodi,
Michael Saxon, Elissa M Redmiles, Sarah Adel Bargal
Newer work: + Eric Zeng

Work in Progress

IV. Defining AI Safety: Mitigations & Gaps

Using prompt inversion

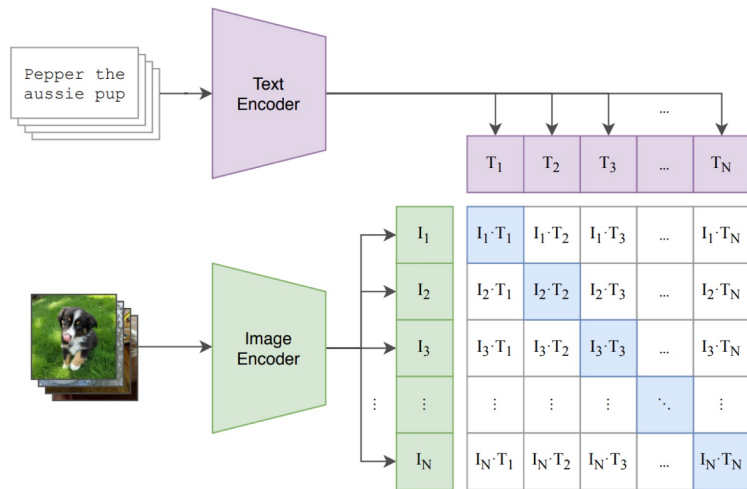
to upper bound risk of SNCL generation from text



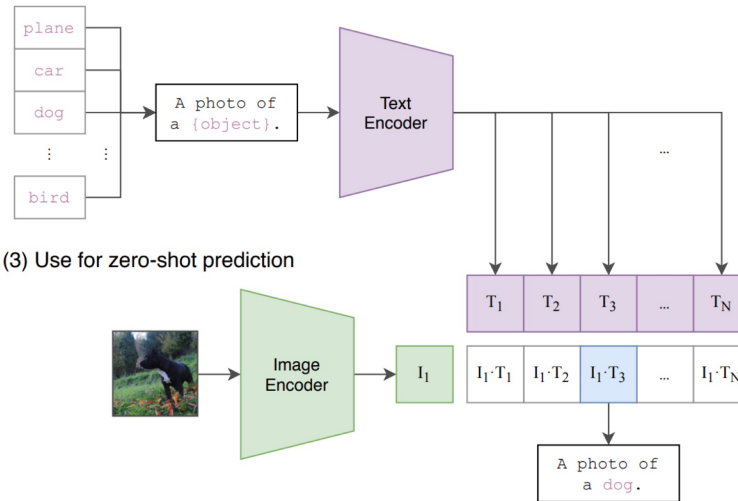
IV. Defining AI Safety: Mitigations & Gaps

Intuition: Finding a shared feature space

(1) Contrastive pre-training



(2) Create dataset classifier from label text



(3) Use for zero-shot prediction

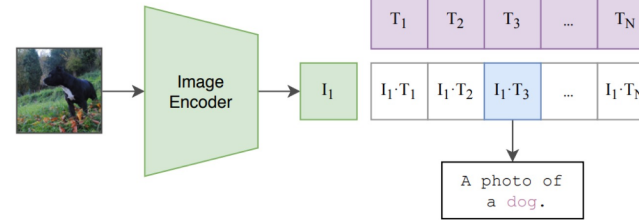
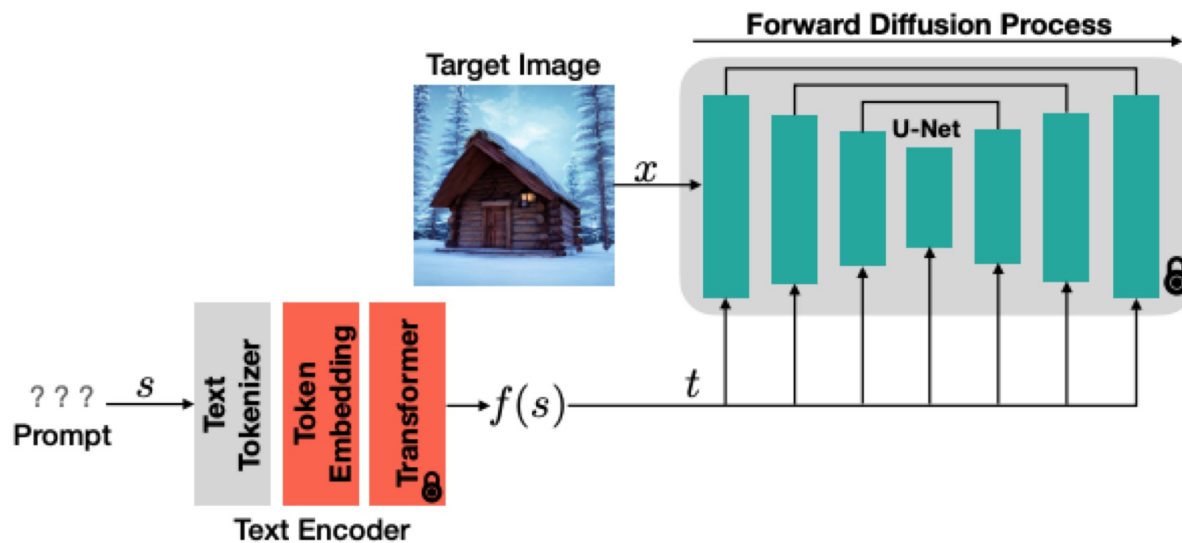


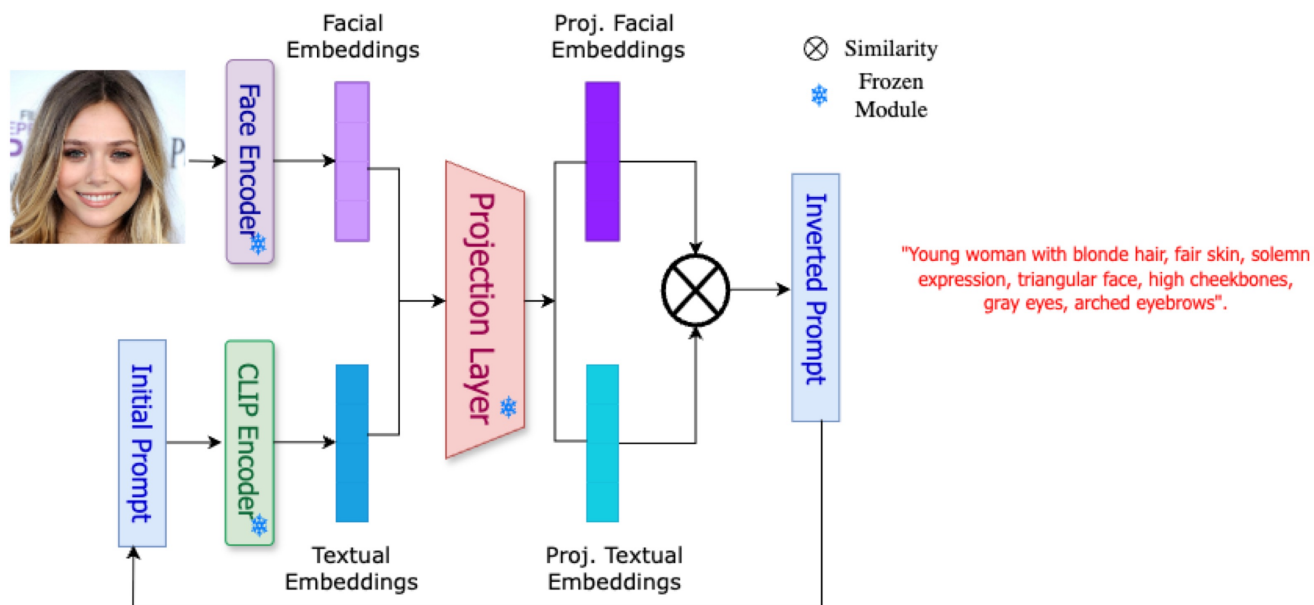
Figure 1. Summary of our approach. While standard image models jointly train an image feature extractor and a linear classifier to predict some label, CLIP jointly trains an image encoder and a text encoder to predict the correct pairings of a batch of (image, text) training examples. At test time the learned text encoder synthesizes a zero-shot linear classifier by embedding the names or descriptions of the target dataset's classes.

Radford et al, Learning Transferable Visual Models From Natural Language Supervision, ICML' 21

Textual Inversion Intuition: optimized search of shared feature space

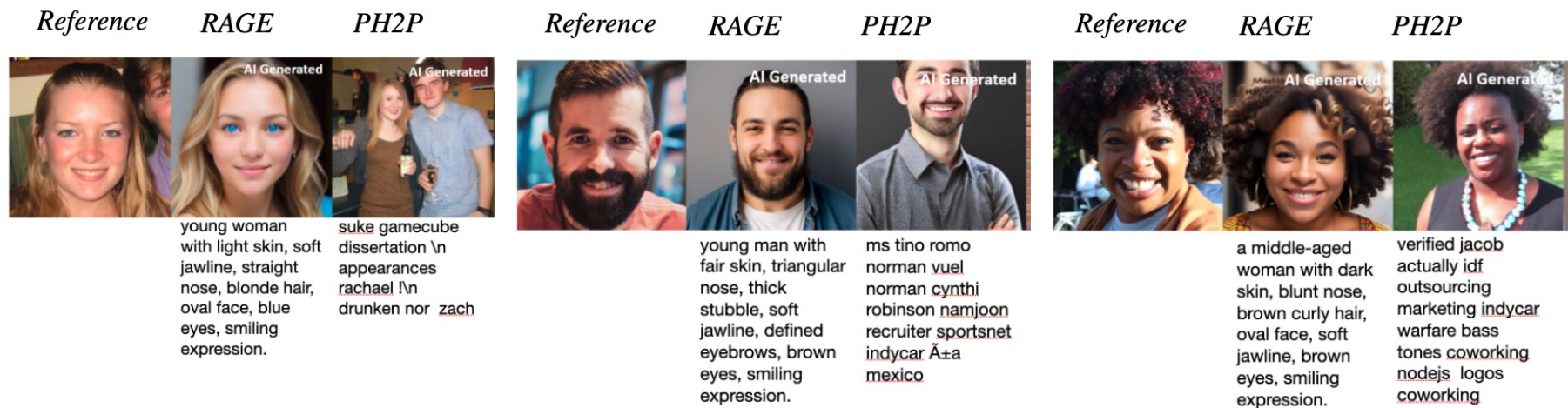


Our RAGE approach uses this intuition with a facial feature dictionary & optimization adjustments



Currently: Use IPAdapter to Generate Images

Future: Test Transferability to Unadapted Models



IV. Defining AI Safety: Mitigations & Gaps

Defining AI Safety

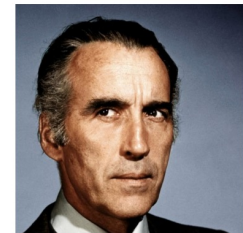
Recognizability:

At what point is an image an image of you?

Recognizability to self vs. close contact vs. acquaintance vs. stranger

How does context change the threshold?

Scale adapted from Phillips et al. PNAS 2018



Consider the two images above. Based on your observations, which of the following would you say is true:

- 3 I am very confident these two images depict the same person.
- 2 I am confident these two images depict the same person.
- 1 I am somewhat confident these two images depict the same person.
- 0 I cannot say whether it is the same person or different persons depicted in the image
- 1 I am somewhat confident these two images do not depict the same person.
- 2 I am confident these two images do not depict the same person.
- 3 I am very confident these two images do not depict the same person.

Finish and Go to Next Image

IV. Defining AI Safety: Mitigations & Gaps

Defining AI Safety

Recognizability:

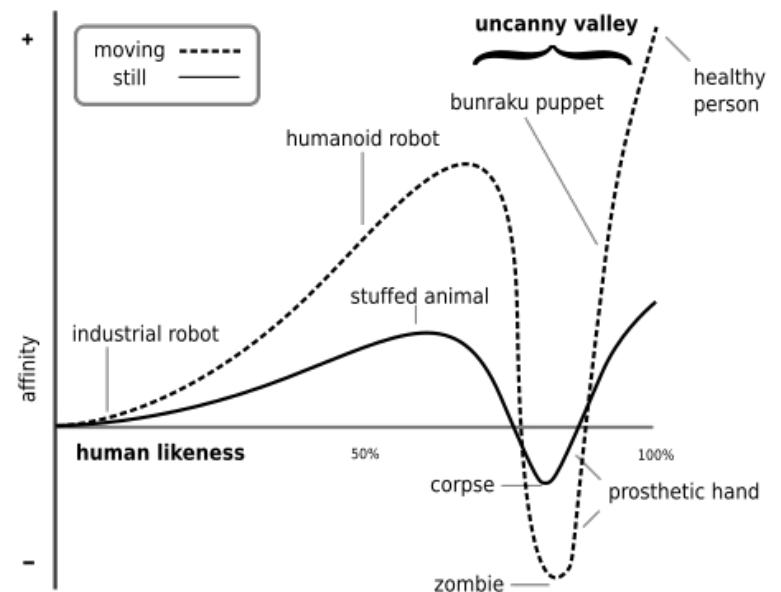
At what point is an image an image of you?

Recognizability to self vs. close contact vs. acquaintance vs. stranger

How does context change the threshold?

How uncanny or unrealistic can the image be?

Masahiro Mori's Uncanny Valley



Example scales for measuring uncanniness

Ordinary (1) 2 3 4 5 6 (7) Creepy

Crude (1) 2 3 4 5 6 (7) Polished

Synthetic (1) 2 3 4 5 6 (7) Real

IV. Defining AI Safety: Mitigations & Gaps

Recognizability

"people can **“file a likeness claim”** that will be reviewed and removed in 24 hours.”

– 404 Media quoting
CivitAI CEO Justin Maier

Defense-in-Depth: Value Chain

“This change is a requirement to continue conversations with specialist payment partners and has to be completed this week to prepare for their service.”

– 404 Media quoting
CivitAI CEO Justin Maier

Defining AI Safety

Safety Theater ~~vs. Threat Dramatics:~~

Can we really do this effectively?

False positives

IV. Defining AI Safety: Mitigations & Gaps

2 Technical Mitigations: AI Generation of NCII

	Mitigation	Assessment
Post-hoc	Perceptual Hashing	✗ Useful for known content, cannot identify new content; increasing AI threat to security of hashes
	Output filters	● Promising; key technical, privacy & free speech challenges remain
	Prompt filters	● Cat-and-mouse but can be friction for UI-bound /script kiddie attackers
	Watermarking / Authenticity	✗ Indicating something is fake or not doesn't stop the harm ● However, watermarks can be used for attribution and watermarks added by generators could be used to identify content & perpetrators
Pre-hoc	Image Poison	● Cat-and-mouse but may add useful friction for “typical” attackers & tools
	Training data filtering	● Important to remove CSAM & NCII from training data but not likely to mitigate generation effectively

Redmiles et al. 2024 <https://mdi.georgetown.edu/formal-response/formal-response-mdi-comment-on-nist-ai-100-4-on-reducing-risks-posed-by-synthetic-content/>
 Redmiles et al. 2024 <https://mdi.georgetown.edu/news/formal-response-mdi-comment-on-nist-ai-executive-order-feb2024/>

IV. Defining AI Safety: Mitigations & Gaps

2

Generative Models

*"We really want to get to a place where we can enable NSFW stuff (e.g. text erotica, gore) for your personal use in most cases...
but not do stuff like make deepfakes."*
-- Sam Altman OpenAI

Pre-hoc defenses (other than image poison)

Vision community: **Model immunization**

Advocacy groups like All Tech Is Human propose as a solution **training data filtering**

IV. Defining AI Safety: Mitigations & Gaps

2

Generative Models

"We really want to get to a place where we can enable NSFW stuff (e.g. text erotica, gore) for your personal use in most cases... but not do stuff like make deepfakes."
-- Sam Altman OpenAI

Pre-hoc defenses (other than image poison)

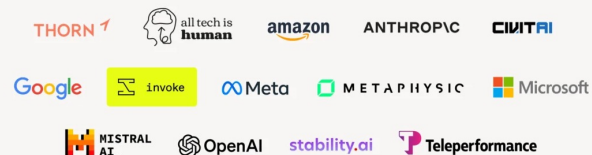
Vision community: **Model immunization**

Advocacy groups like All Tech Is Human propose as a solution **training data filtering**

Thorn and All Tech Is Human Forge
Generative AI Principles with AI Leaders to
Enact Strong Child Safety Commitments

July 16, 2024 — 11 Minute Read

SHARE   



IV. Defining AI Safety: Mitigations & Gaps

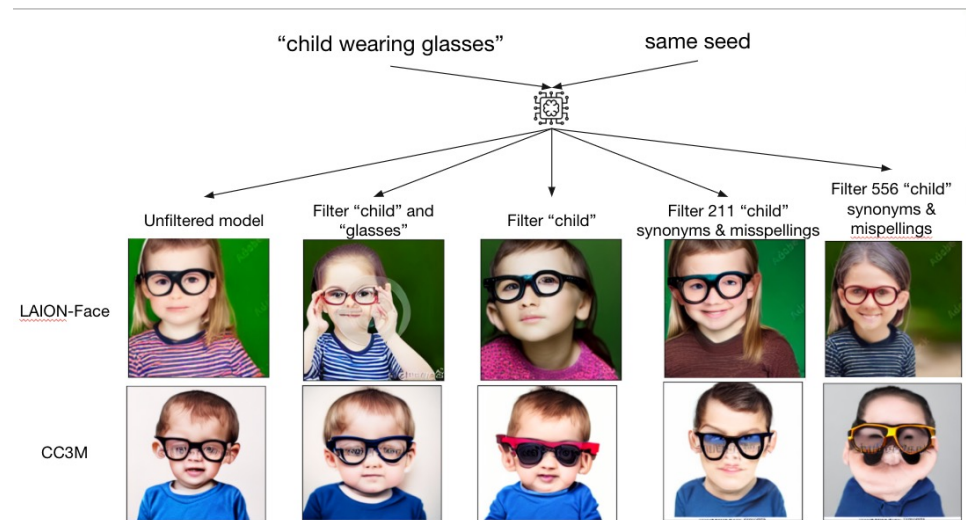
2

Generative Models

"We really want to get to a place where we can enable NSFW stuff (e.g. text erotica, gore) for your personal use in most cases... but not do stuff like make deepfakes."
 -- Sam Altman OpenAI

Pre-hoc defenses (other than image poison)

training data filtering



Work with Ana-Maria Crețu, Klim Kireev, Wisdom Obinna, Amro Abdalla, Raphael Meier, Sarah Bargal, Elissa Redmiles, Carmela Troncoso

IV. Defining AI Safety: Mitigations & Gaps

2

Generative Models

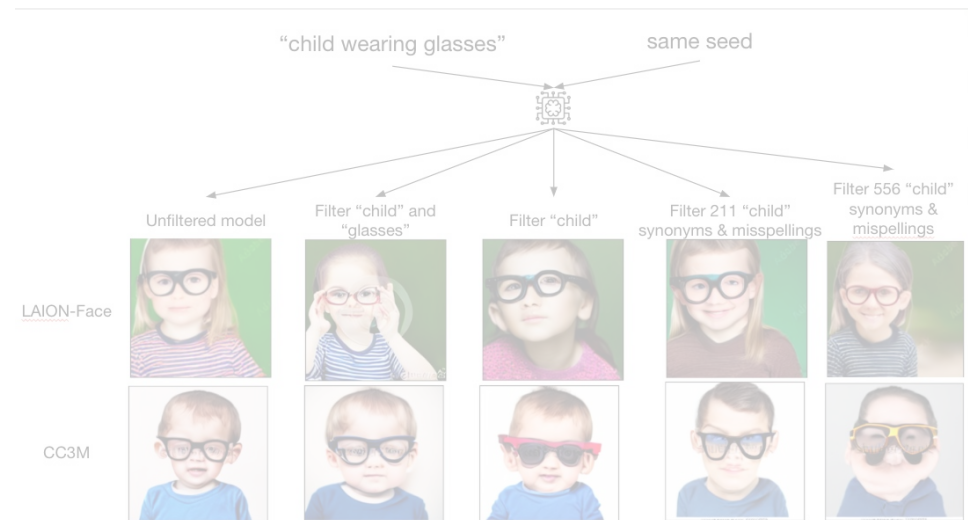
*"We really want to get to a place where we can enable NSFW stuff (e.g. text erotica, gore) for your personal use in most cases...
but not do stuff like make deepfakes."*
 -- Sam Altman OpenAI

Defining AI Safety

How uncanny or unrealistic can the image be?

Pre-hoc defenses (other than image poison)

training data filtering



Work with Ana-Maria Crețu, Klim Kireev, Wisdom Obinna, Amro Abdalla, Raphael Meier, Sarah Bargal, Elissa Redmiles, Carmela Troncoso

IV. Defining AI Safety: Mitigations & Gaps

1

“Nudification” applications

These websites and applications purport to use computer vision techniques to undress the subject of an uploaded image

2

Generative Models

People may use generative models **hosted online** to generate SNCII of an individual by e.g., inputting a URL to an image of them as part of their prompt or they may run and fine tune their own **local models**

3

Hiring experts: Face swap + custom toolkits

Perpetrators may also hire experts who use a combination of face swapping, generative models, and image editing tools to create SNCII of a target.

Recent research on *videos* discovers 161 GitHub repositories, 60 of which link to CS papers (Han, Li, Kumar & Durumeric 2024).

IV.

(Mis)use of Nude Images in Machine Learning Research

Arshia Arya³ Princessa Cintaqia² Deepak Kumar³
Allison McDonald² Lucy Qin¹ Elissa M. Redmiles¹
¹Georgetown University ²Boston University ³University of California, San Diego
{lucy.qin, elissa.redmiles}@georgetown.edu
{cintaqia, amcdon}@bu.edu {aarshia, kumarde}@ucsd.edu

1 Introduction

Nudity detection is a task that has been studied by researchers for decades [1]. For training, testing, and benchmarking nudity detection algorithms, researchers typically scrape images from the internet or use existing datasets of nude images. While this practice is common for assembling datasets for general image-recognition tasks, nude images are particularly sensitive. Images that were consensually shared on publicly-accessible forums (e.g., Reddit) or adult content platforms (e.g., Pornhub, OnlyFans) were never intended to be used in research. Furthermore, publicly-accessible forums have been documented to host communities explicitly for the nonconsensual sharing of nude content [2]. Such sharing is a common form of image-based sexual abuse (IBSA) [3], which is a category of technology-facilitated sexual violence that includes the nonconsensual creation and distribution of intimate content. IBSA can lead to serious legal, emotional, employment, and relational consequences [4, 5], including clinical diagnoses of post-traumatic stress disorder, anxiety, and/or depression [5]. One of the most traumatizing aspects is that once an image has been distributed online, people lose control over how it is further spread and used. A victim-survivor shared with Bates et al. that, “I didn’t have control over who they were distributed to [...] that they were used maliciously and without my consent, and in my name, that was the part that violated me the most” [5].

Our team is currently conducting research that investigates the use of nude datasets in machine learning and computer vision literature. By keyword searching for common ML tasks involving nudity (see Table A), we found a seed set of 2048 papers. While we are still processing the full results, we identify several ethical challenges based on our initial observations after reading dozens of these papers. In this provocation, we aim to raise questions for researchers considering work in this space to evaluate at the start of their projects and prior to dataset collection.

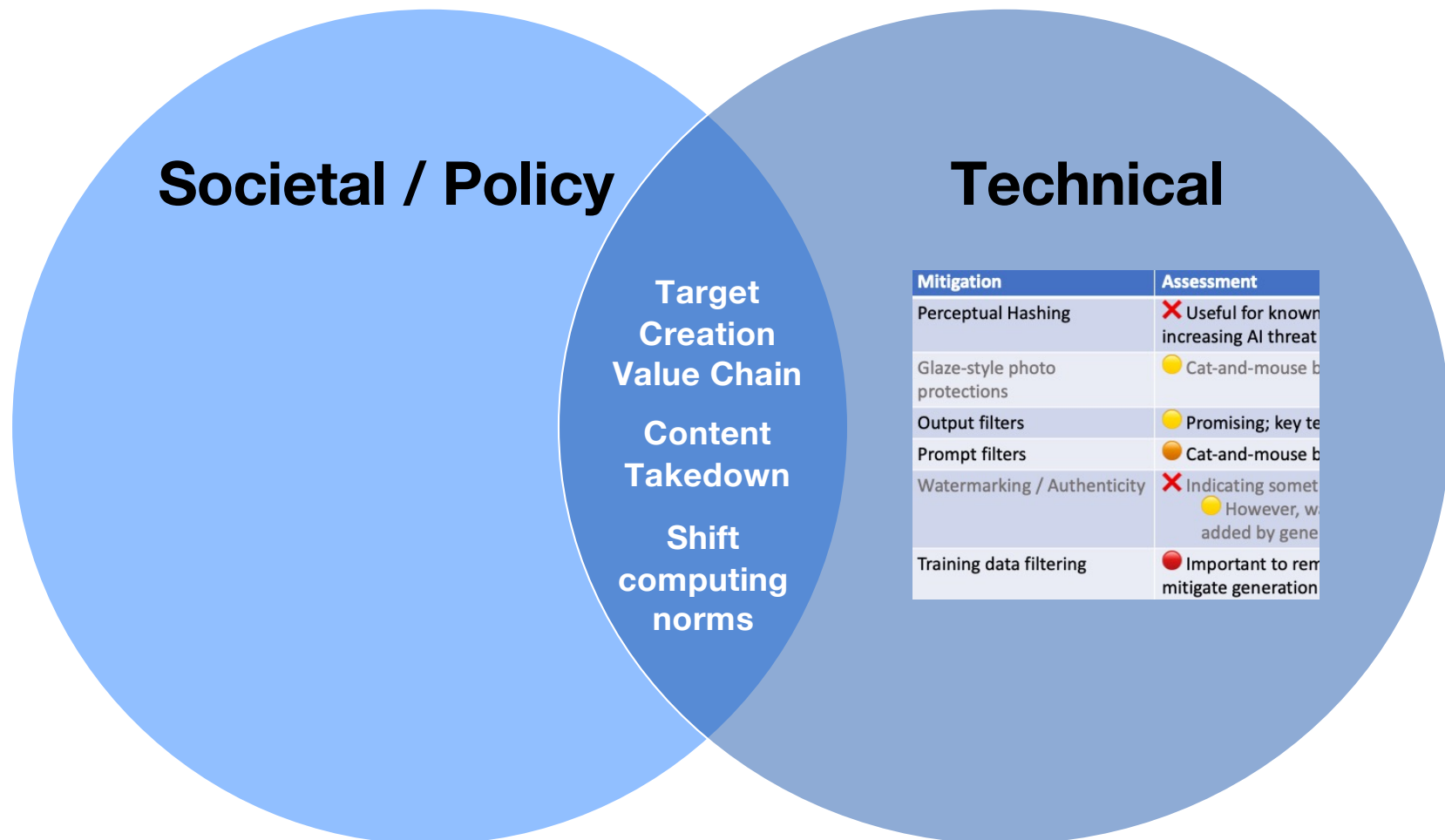
3

Hiring experts: Face swap + custom toolkits

Perpetrators may also hire experts who use a combination of face swapping, generative models, and image editing tools to create SNCH of a target.

Recent research on *videos* discovers 161 GitHub repositories, 60 of which link to CS papers (Han, Li, Kumar & Durumeric 2024).

IV. Defining AI Safety: Mitigations & Gaps



“Violation of my body”

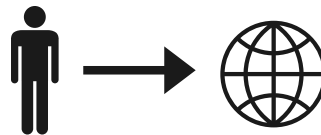
Perception of AI-generated non-consensual (intimate) imagery

Natalie Grace Brigham, Miranda Wei,
Tadayoshi Kohno, Elissa M Redmiles

USENIX SOUPS
2024

III. SNCII: AI Generation

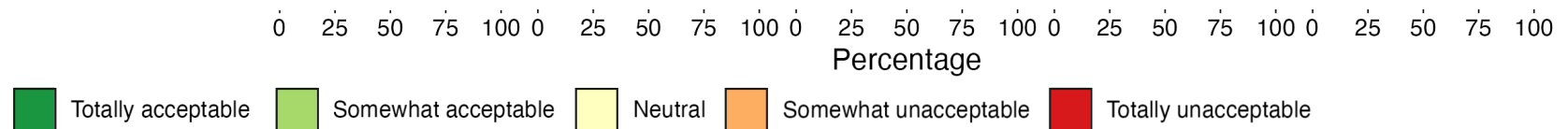
A stranger uses AI to
create a synthetic video of
you **performing a sexual
act** and posts it publicly



III. SNCII: AI Generation

Key finding

- V11 - stranger/sexual act/harm
- V2 - intimate partner/sexual act/harm
- V12 - stranger/sexual act/sexual pleasure
- V10 - stranger/sexual act/entertainment
- V1 - intimate partner/sexual act/entertainment
- V3 - intimate partner/sexual act/sexual pleasure



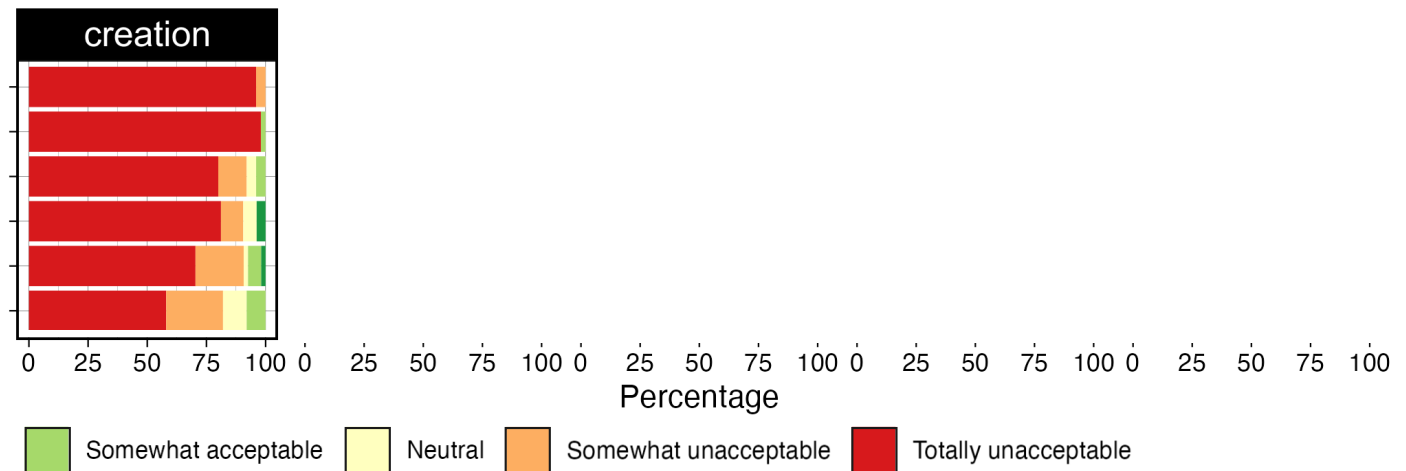
Respondents' perceptions of acceptability across vignettes involving SNCII; each vignette is defined by the creator / action / intent.

Survey, December 2023 (n=315); census-representative sample

III. SNCII: AI Generation

Key finding

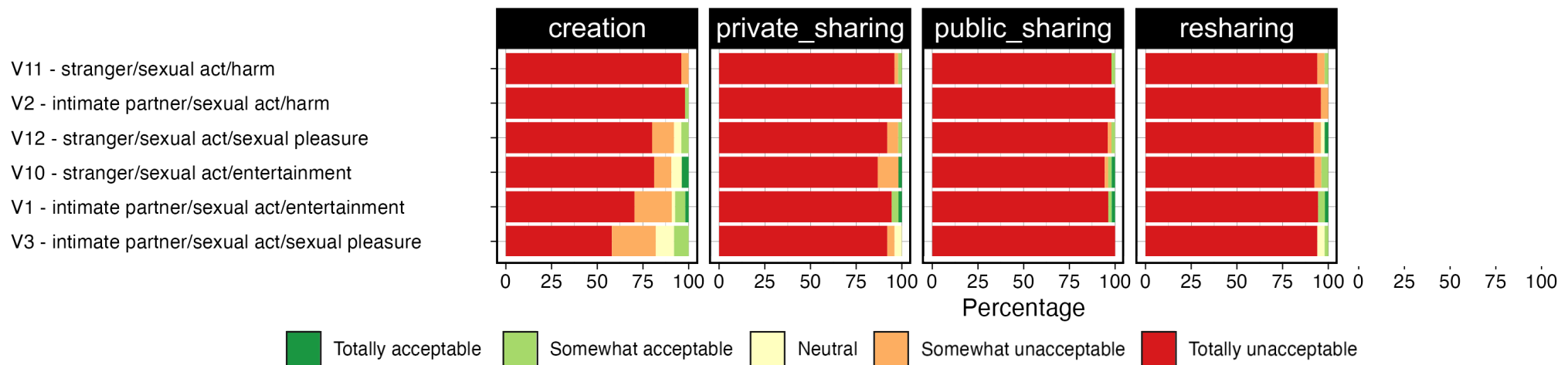
- V11 - stranger/sexual act/harm
- V2 - intimate partner/sexual act/harm
- V12 - stranger/sexual act/sexual pleasure
- V10 - stranger/sexual act/entertainment
- V1 - intimate partner/sexual act/entertainment
- V3 - intimate partner/sexual act/sexual pleasure



Respondents' perceptions of acceptability across vignettes involving SNCII; each vignette is defined by the creator / action / intent.

III. SNCII: AI Generation

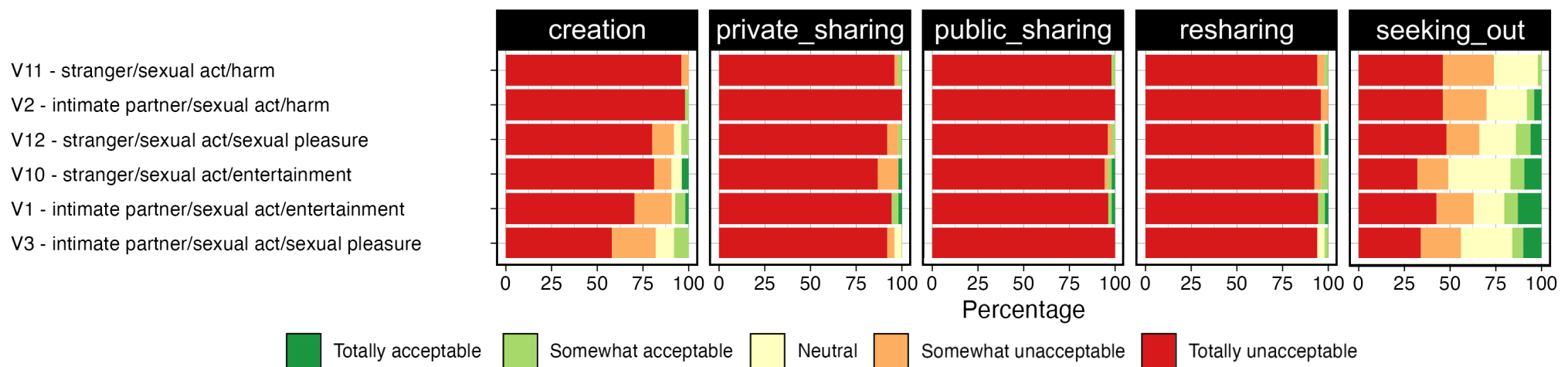
Key finding



Respondents' perceptions of acceptability across vignettes involving SNCII; each vignette is defined by the creator / action / intent.

III. SNCII: AI Generation

Viewing is perceived as far more acceptable



Respondents' perceptions of acceptability across vignettes involving SNCII; each vignette is defined by the creator / action / intent.

Viewing is perceived as far more acceptable

90%+

perceive **creating & sharing** SNCII as unacceptable

Barely

50%

perceive **seeking it out** to view as unacceptable

III. SNCII: AI Generation

It's hard to measure abuse prevalence but...

Recent surveys (n>1600, USA) suggest that

~0.7% of Americans have **created** SNCII

~7% have **seen** SNCII

Defense-in-Depth

TARGET VIEWING NORMS

Viewing may be perceived as acceptable due to a sense of **distributed responsibility**: ethical responsibility is spread across the network of online actors (Ess 2014, de Vries 2022)

Deterrence messages can be used to break this sense of distributed responsibility

IV. Defining AI Safety: Mitigations & Gaps

Defense-in-Depth **TARGET VIEWING NORMS**

1) Defining AI Safety - **Dual Use:** Create pipelines that identify misuse using contextual signals (ad language, search terms)

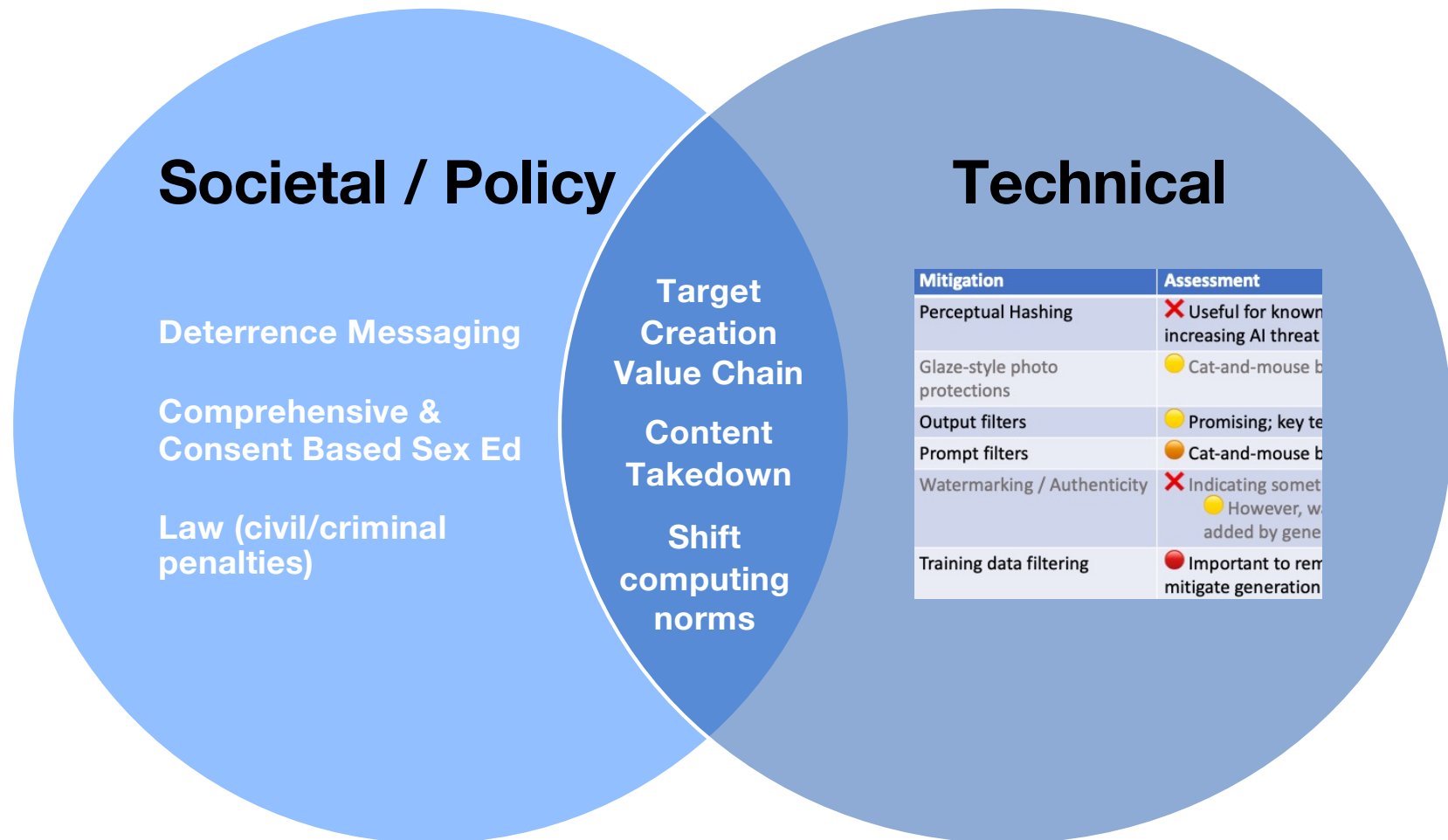
2) Map those contextual signals to psychological research on perpetrator motivations

3) Use them to target personalized, just-in-time deterrence messages



Work with Asia Eaton (Psychology)
Amy Hasinoff (Communications)
Yoshi Kohno & Eric Zeng (Ads)

IV. Defining AI Safety: Mitigations & Gaps



IV. Defining AI Safety: Mitigations & Gaps



Innovate Technology

Need: defense-in-depth to add friction to distribution & creation

Need: scalable AI safety risk assessment instruments

IV. Defining AI Safety: Mitigations & Gaps



Innovate Technology

Need: defense-in-depth to add friction to distribution & creation

Need: scalable AI safety risk assessment instruments



Shape Norms

Need: consent-based sex education via UI (e.g., deterrence messaging) & classroom interventions

Need: ethical benchmarking datasets in our community: don't perpetrate the harms we aim to stop

IV. Defining AI Safety: Mitigations & Gaps



Innovate Technology

Need: defense-in-depth to add friction to distribution & creation

Need: scalable AI safety risk assessment instruments



Shape Norms

Need: consent-based sex education via UI (e.g., deterrence messaging) & classroom interventions

Need: ethical benchmarking datasets in our community: don't perpetrate the harms we aim to stop



Implement Features & Policy

Need: to deploy friction features consistently across platforms

Need: to deploy & enforce policy that penalizes perpetration



Innovate Technology

Need: defense-in-depth to add friction to distribution & creation

Need: scalable AI safety risk assessment instruments



Shape Norms

Need: consent-based sex education via UI (e.g., deterrence messaging) & classroom interventions

Need: ethical benchmarking datasets in our community: don't perpetrate the harms we aim to stop



Implement Features & Policy

Need: to deploy friction features consistently across platforms

Need: to deploy & enforce policy that penalizes perpetration

Safe(r) Digital Intimacy Session II: Sex, Work, and Technology + Ideation Session

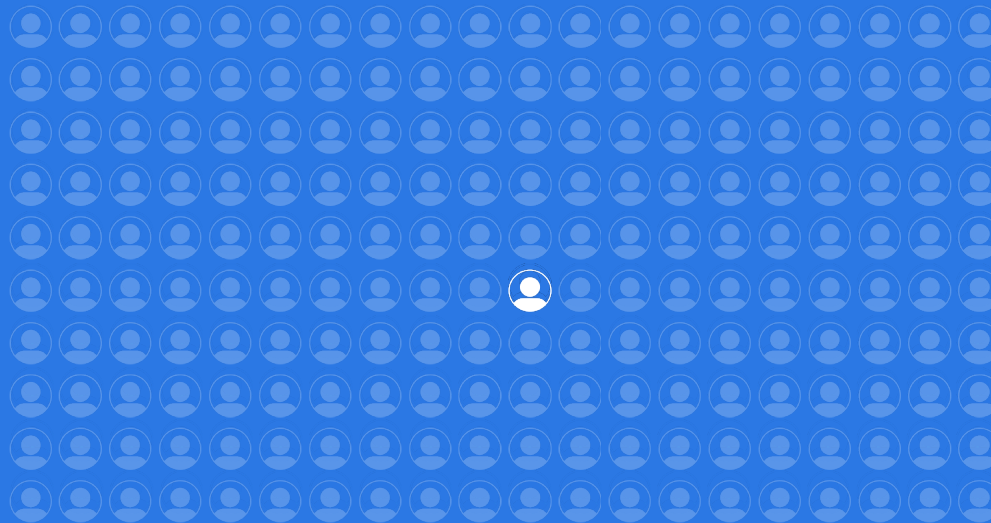
Elissa M. Redmiles

Clare Luce Boothe Assistant Professor of Computer Science
Georgetown University

elissa.redmiles@georgetown.edu

I. CONTEXT SETTING

The United Nations estimates that



as many as

1 in 200

people have worked in the
sex industry* in their lifetime

I. CONTEXT SETTING

Technology is used in sex work* in many different ways



In-Person Sex Work

Digital-Only Sex Work

* Sex work is defined as the exchange of sexual services for money, encompassing a broad range of services such as escorting (i.e., full-service sex work), erotic massage, porn acting, camming (performing live sex acts on video), phone sex, professional domination (performing the dominant role in a BDSM relationship), and erotic dancing

I. CONTEXT SETTING

Sex work is situated in a complex legal landscape & is highly stigmatized

Full service sex work is legal or decriminalized in Belgium, Germany, Switzerland, UK & more

Sexual content sale is legal in more places including the U.S.

agenda

I. Context Setting

II. Challenges Facing Digitally-Mediated Sex Work

III. Growth of Professional Sexual Content Creation

IV. Opportunities: Engineering & Internet Governance

V. Ideation Time!

II. CHALLENGES FACING DIGITALLY-MEDIATED SEX WORK

STARTING POINT:

What **challenges** do **in-person sex workers** face with **digital mediation**?

II. CHALLENGES FACING DIGITALLY-MEDIATED SEX WORK

Two-year mixed methods research project

Barwulor, C., McDonald, A., Hargittai, E., and Redmiles, E.M.
"Disadvantaged in the American-dominated internet": Sex, Work, and Technology. ACM CHI 2021.

McDonald, A., Barwulor, C., Mazurek, M.L., Schaub, F., and Redmiles, E.M. **"It's stressful having all these phones": Investigating Sex Workers' Safety Goals, Risks, and Practices Online.** USENIX Security Symposium 2021.
Distinguished Paper Award Winner. 

RESEARCH QUESTION 1

How do sex workers (want to) use technology?

RESEARCH QUESTION 2

What digital challenges do sex workers face?

II. CHALLENGES FACING DIGITALLY-MEDIATED SEX WORK

Two-year mixed methods research project

3 Countries where sex work is legal: Germany, Switzerland, and the UK

29 Interviews with sex workers conducted via end-to-end encrypted chat, phone, or video

65 Surveys with sex workers

Participants [recruited through in-person flyers](#), sex work organizations, and snowball sampling

Protocol and study materials were [reviewed by sex working consultant](#) to ensure the appropriateness and ethics of materials.

II. CHALLENGES FACING DIGITALLY-MEDIATED SEX WORK

Two-year
mixed methods
research project

RESEARCH QUESTION 1

**How do sex workers (want to)
use technology?**

RESEARCH QUESTION 2

**What digital challenges do
sex workers face?**

II. CHALLENGES FACING DIGITALLY-MEDIATED SEX WORK

RESEARCH QUESTION 1: HOW DO SEX WORKERS (WANT TO) USE TECHNOLOGY?

Six primary technology needs for sex workers:



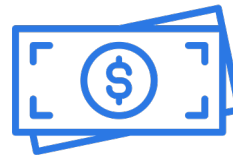
Advertising



Client Vetting



**Client
Communication**



Payments



**Support/
Education**



Activism

II. CHALLENGES FACING DIGITALLY-MEDIATED SEX WORK

RESEARCH QUESTION 1: HOW DO SEX WORKERS (WANT TO) USE TECHNOLOGY?

Six primary technology needs for sex workers:



Advertising



Client Vetting

- Verify client is who they say they are
- Check whether other workers have had negative experiences with them



Client
Communication



Payments



Support/
Education



Activism

II. CHALLENGES FACING DIGITALLY-MEDIATED SEX WORK

RESEARCH QUESTION 1: HOW DO SEX WORKERS (WANT TO) USE TECHNOLOGY?

Six primary technology needs for sex workers:



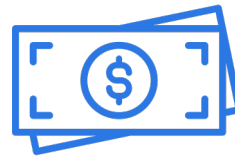
Advertising



Client Vetting



Client
Communication



Payments

- Collect digital (i.e., remote) payment (Venmo, Paypal, CashApp)
- Receive gifts and gift cards



Support/
Education



Activism

II. CHALLENGES FACING DIGITALLY-MEDIATED SEX WORK

RESEARCH QUESTION 1: HOW DO SEX WORKERS (WANT TO) USE TECHNOLOGY?

Six primary technology needs for sex workers:



Advertising



Client Vetting



Client
Communication



Payments



**Support/
Education**



Activism

- Connect with other sex workers for support and community
- Share information about bad clients
- Share tips on staying safe

II. CHALLENGES FACING DIGITALLY-MEDIATED SEX WORK

Two-year mixed methods research project

RESEARCH QUESTION 1

How do sex workers (want to) use technology?

RESEARCH QUESTION 2

What digital challenges do sex workers face?

McDonald, A., Barwulor, C., Mazurek, M.L., Schaub, F., and Redmiles, E.M. “It’s stressful having all these phones”: Investigating Sex Workers’ Safety Goals, Risks, and Practices Online. USENIX Security Symposium 2021. Distinguished Paper Award Winner.

II. CHALLENGES FACING DIGITALLY-MEDIATED SEX WORK

RESEARCH QUESTION 2: WHAT DIGITAL CHALLENGES DO SEX WORKERS FACE?

Platform Censorship

Platforms censor sexual expression and sex work, regardless of the legality of that expression or labor

II. CHALLENGES FACING DIGITALLY-MEDIATED SEX WORK

RESEARCH QUESTION 2: WHAT DIGITAL CHALLENGES DO SEX WORKERS FACE?

Platform Censorship



II. CHALLENGES FACING DIGITALLY-MEDIATED SEX WORK

RESEARCH QUESTION 2: WHAT DIGITAL CHALLENGES DO SEX WORKERS FACE?

Platform Censorship



“I had, I don’t know how many followers on Instagram and at some point, I hadn’t even posted [nudes] or something, at some point it was just deleted... So that definitely hurt my business, but not in a way that I bled to death or something. But that was pretty shitty.”

Research Participant

II. CHALLENGES FACING DIGITALLY-MEDIATED SEX WORK

RESEARCH QUESTION 2: WHAT DIGITAL CHALLENGES DO SEX WORKERS FACE?

Platform Censorship



“As someone who offers proactive erotic services, you are clearly at a disadvantage in the American-dominated internet. There is censorship (content that must not be present, page blocks, photos that must not be shown etc.) and restrictions...”

Research Participant

II. CHALLENGES FACING DIGITALLY-MEDIATED SEX WORK

RESEARCH QUESTION 2: WHAT DIGITAL CHALLENGES DO SEX WORKERS FACE?

Platform Censorship



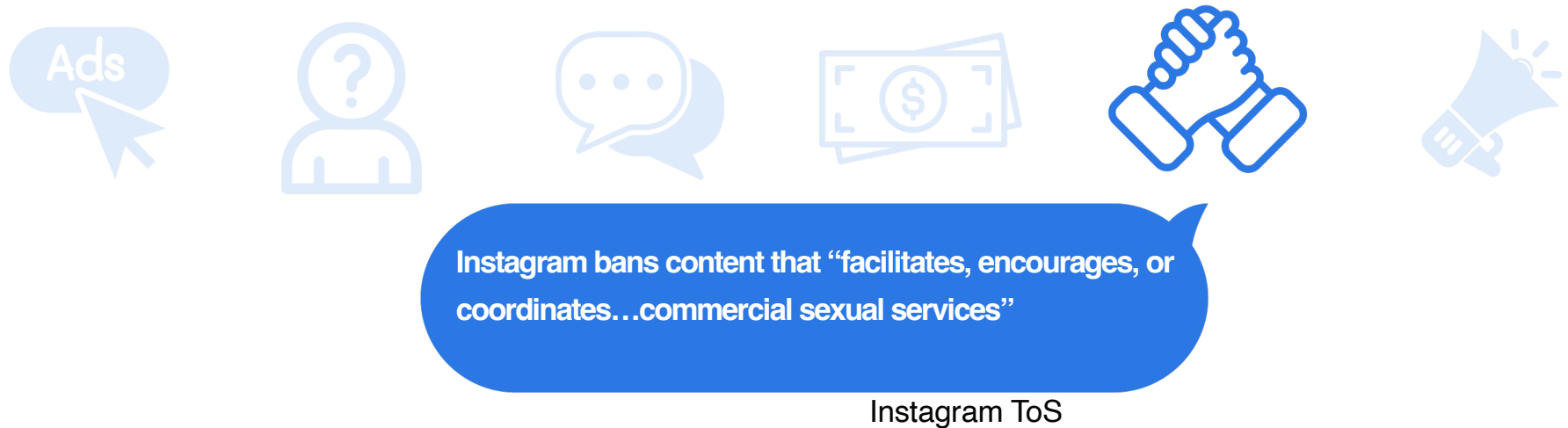
“PayPal would be great, but sex work is forbidden there and I have colleagues who did that and they got blocked for life by PayPal...”

Research Participant

II. CHALLENGES FACING DIGITALLY-MEDIATED SEX WORK

RESEARCH QUESTION 2: WHAT DIGITAL CHALLENGES DO SEX WORKERS FACE?

Platform Censorship



II. CHALLENGES FACING DIGITALLY-MEDIATED SEX WORK

RESEARCH QUESTION 2: WHAT DIGITAL CHALLENGES DO SEX WORKERS FACE?

Technology platforms a
regulate sex worker's identity,
not just their work

Sex Workers Have Been Shunned by Banks, Even When Their Work Is Legal

Financial service companies often avoid what they deem high-risk industries like adult entertainment. When workers lose their accounts, they are left with few options.

VICE Video TV The 8:46 Project News World News Tech Music

Identity

Why It's Perfectly Legal for Airbnb to Discriminate Against Sex Workers

A professional dominatrix says Airbnb suspended her account because of her job—even though she wasn't using the app to do sex work.

Sex Workers Took Refuge in Crypto. Now It's Failing Them

Banks and payments companies have long penalized sex workers. Many thought crypto would be a solution, but now exchanges are dumping them too.

II. CHALLENGES FACING DIGITALLY-MEDIATED SEX WORK

RESEARCH QUESTION 2: WHAT DIGITAL CHALLENGES DO SEX WORKERS FACE?

**Technology platforms a
regulate sex worker's identity,
not just their work**

Sex Workers Have Been

“AirBnB bans workers just for being [sex workers]. They have not shown their face, don't use same email or phone... and they don't work from AirBnB and they got banned.”

Research Participant

**Sex Workers Took Refuge in Crypto.
Now It's Failing Them**

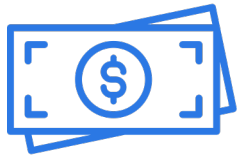
Banks and payments companies have long penalized sex workers. Many thought crypto would be a solution, but now exchanges are dumping them too.

II. CHALLENGES FACING DIGITALLY-MEDIATED SEX WORK

RESEARCH QUESTION 2: WHAT DIGITAL CHALLENGES DO SEX WORKERS FACE?

Platform Censorship

Censorship is a safety problem



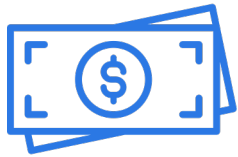
Criminals know sex workers can't use digital payment platforms. They linger in areas with high brothel density to rob workers leaving work with large amounts of cash.

II. CHALLENGES FACING DIGITALLY-MEDIATED SEX WORK

RESEARCH QUESTION 2: WHAT DIGITAL CHALLENGES DO SEX WORKERS FACE?

Platform Censorship

Censorship is a safety problem



Criminals know sex workers can't use digital payment platforms. They linger in areas with high brothel density to rob workers leaving work with large amounts of cash.



Inability to form online community restricts ability to vet clients & limits access to health and safety education from peers.

II. CHALLENGES FACING DIGITALLY-MEDIATED SEX WORK

RESEARCH QUESTION 2: WHAT DIGITAL CHALLENGES DO SEX WORKERS FACE?

Platform Censorship

Outing & Context Collapse

Others discovering sex work or related (e.g., LGBTQ+) identities can lead to violence, loss of housing, etc.

II. CHALLENGES FACING DIGITALLY-MEDIATED SEX WORK

RESEARCH QUESTION 2: WHAT DIGITAL CHALLENGES DO SEX WORKERS FACE?

Platform Censorship

Outing & Context Collapse

Clients can be aggressive and/or invasive

“My lovely partner, who is also a photographer, has photographed me a couple times. I wasn’t very smart and published my photos with his tag on a relatively public forum. . . Then, a client of mine who was very fond of me— which I also wasn’t totally aware of— did some research and figured out who my boyfriend is.

He found our places of business and then of course knew what we do in our free time, what our names are... Since then, I pay extremely close attention which tag is on the pictures.”

Research Participant

II. CHALLENGES FACING DIGITALLY-MEDIATED SEX WORK

RESEARCH QUESTION 2: WHAT DIGITAL CHALLENGES DO SEX WORKERS FACE?

Platform Censorship

Outing & Context Collapse

Laws regulating sex work and business
can create exposure risks

II. CHALLENGES FACING DIGITALLY-MEDIATED SEX WORK

Two-year mixed methods research project

McDonald, A., Barwulor, C., Mazurek, M.L., Schaub, F., and Redmiles, E.M. “It’s stressful having all these phones”: Investigating Sex Workers’ Safety Goals, Risks, and Practices Online. USENIX Security Symposium 2021. Distinguished Paper Award Winner.

RESEARCH QUESTION 1

How do sex workers (want to) use technology?

RESEARCH QUESTION 2

What digital challenges do sex workers face?

STRATEGIES

What digital strategies do sex workers use to stay safe?

WHAT STRATEGIES DO SEX WORKERS USE TO STAY SAFE?

Safety Strategies: **During-Service Violence**

68% Covering

giving a friend or colleague appointment details
and checking in after an appointment

51% Vetting Clients

verifying a client is who they say they are and checking
their reputation among other workers

WHAT STRATEGIES DO SEX WORKERS USE TO STAY SAFE?

Safety Strategies: **Preventing Censorship, Outing & Context Collapse**

Identity Management

77% **Alias**

66% **Separate Accounts/Devices**

“It’s also time-consuming, and it’s annoying.... It’s stressful having all these phones and personas and things I have to remember. I’m like, ‘Shit, did I miss that when I put this up?’ All the time.”

Research Participant

WHAT STRATEGIES DO SEX WORKERS USE TO STAY SAFE?

Safety Strategies: **Preventing Censorship, Outing & Context Collapse**

46% Self Censorship

“I am legally allowed to work.... [However,] I am scared of getting banned from certain countries just for being a sex worker so I remove all my info, account and website and wipe my phone before travelling.”

Research Participant

WHAT STRATEGIES DO SEX WORKERS USE TO STAY SAFE?

(Not so useful) Safety Strategies: **Security Tools**

32%

Encrypted Chat

14%

VPN

9%

Encrypted Email

8%

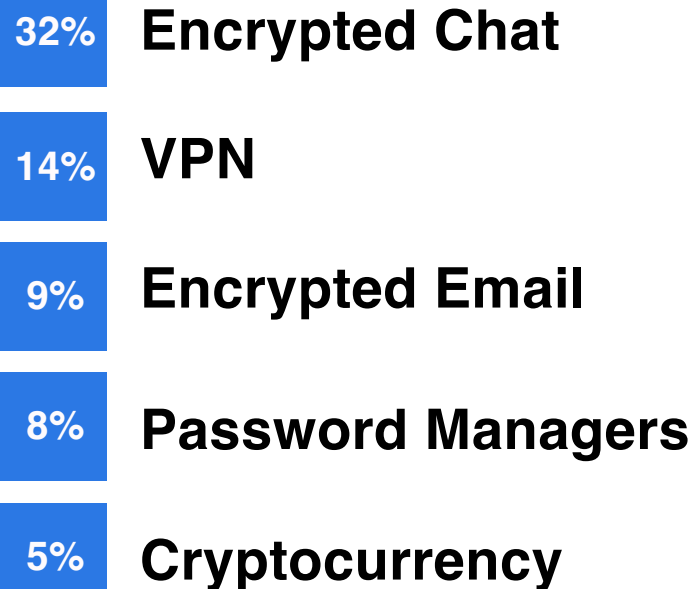
Password Managers

5%

Cryptocurrency

WHAT STRATEGIES DO SEX WORKERS USE TO STAY SAFE?

(Not so useful) Safety Strategies: **Security Tools**



“I would gladly do all of the above, but that really only works when the customers participate: Threema / Signal / Telegram, PGP-encryption, cryptocurrency...”

Research Participant

WHAT STRATEGIES DO SEX WORKERS USE TO STAY SAFE?

(Not so trustworthy) Safety Strategies: **Security Settings**

35% Security and Privacy Settings

“I think it doesn’t matter if you put [settings] on or off...”

Research Participant

II. CHALLENGES FACING DIGITALLY-MEDIATED SEX WORK

WHAT STRATEGIES DO SEX WORKERS USE TO STAY SAFE?

Necessity of relying on manual strategies → **Resignation & Regret**

Managing separate accounts, identities, devices, and information is hard, and mistakes feel inevitable.

Undoing past mistakes is hard.

“I login to Kaufmich in browser on my personal phone and my Apple account for work phone is registered to my passport name... This stuff could get me killed or deported.... I am not prepared.”

Research Participant

II. CHALLENGES FACING DIGITALLY-MEDIATED SEX WORK

Summary

RESEARCH QUESTION 1

How do sex workers (want to) use technology?



RESEARCH QUESTION 2

What digital (safety) challenges do they face?



Platform
Censorship



Outing &
Context Collapse

STRATEGIES

What digital strategies do they use to stay safe?



Identity
Management



Self Censorship



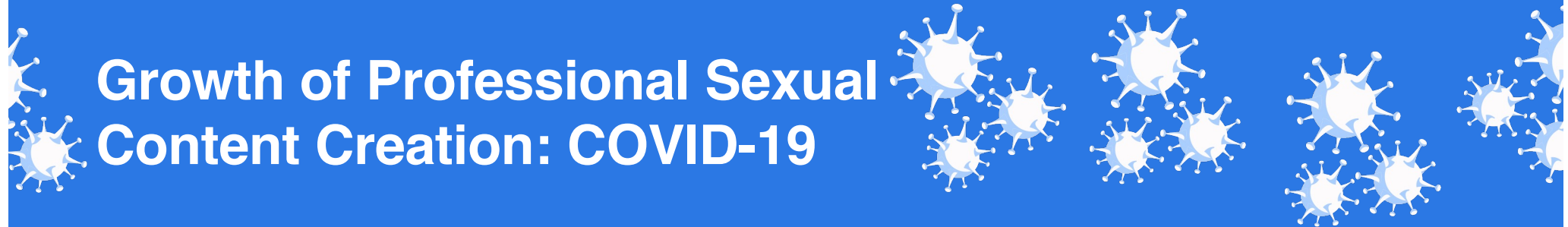
Security Tools



Fail to address threat
model + power
dynamics

III.

Growth of Professional Sexual Content Creation: COVID-19



III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

Existing sex workers **couldn't do in-person work** for two primary reasons



**Risk for Sex
Workers &
Clients**



**Lockdown Laws
& Regulations**

Remember from Yesterday: NDII: Intimate Sharing

Participants used 40+ platforms

Any platform that can create, share or store visual content **is likely used for intimate content**



Messaging
(apps and SMS)



**Social Media
Platforms**



**Hookup/
Dating
Apps**



**Adult Content
Platforms**



**File Share
Platforms**



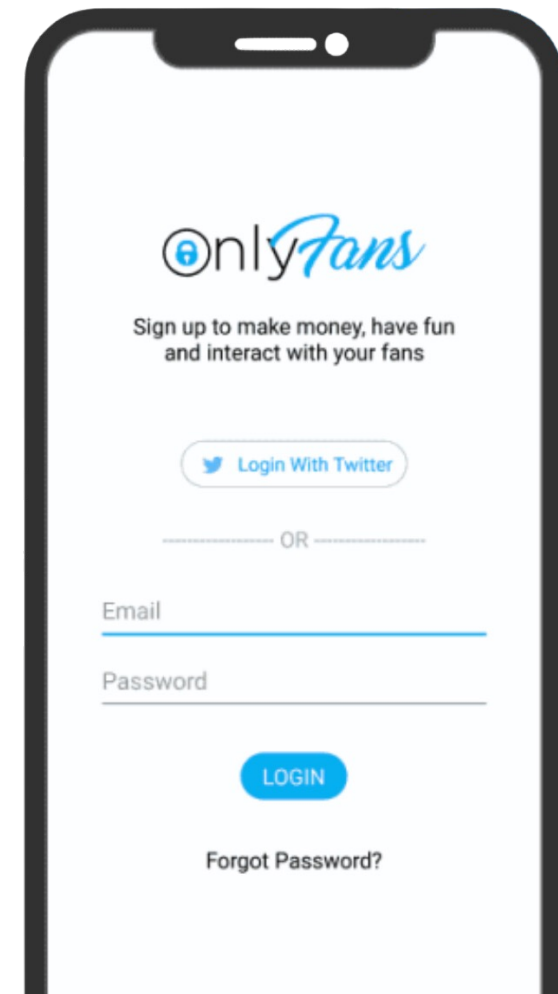
Email

III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

OnlyFans

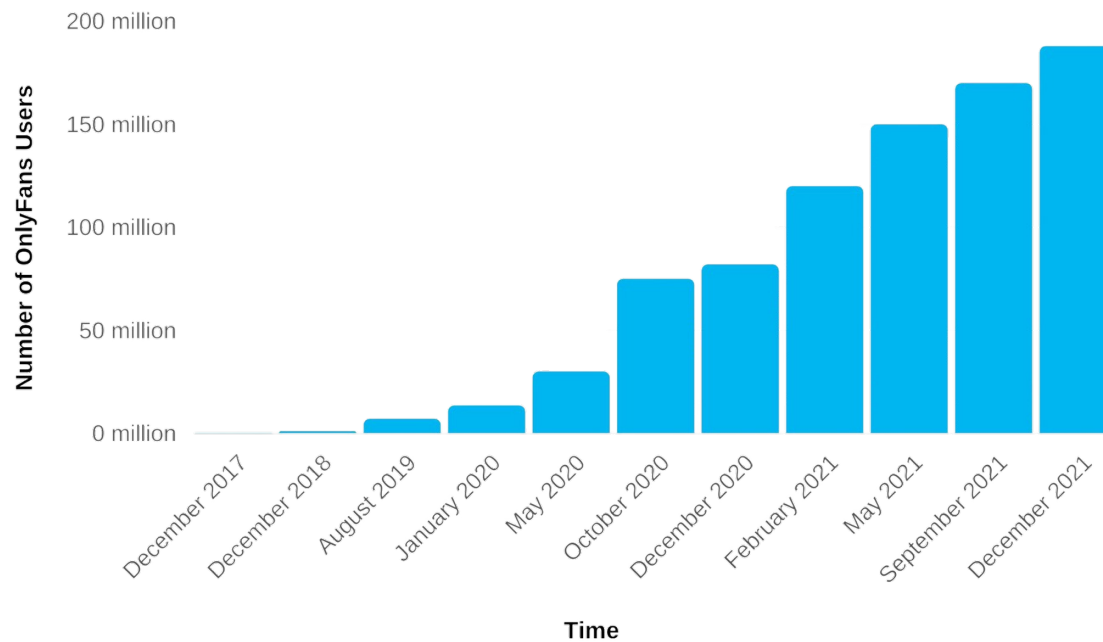
started in 2016 by two people from the webcam industry

It uses a “digital patronage” model, where subscribers can buy a subscription to a feed (where they see content), message creators (often paid), request custom content (paid), pay tips, buy premium pay-per-view content.



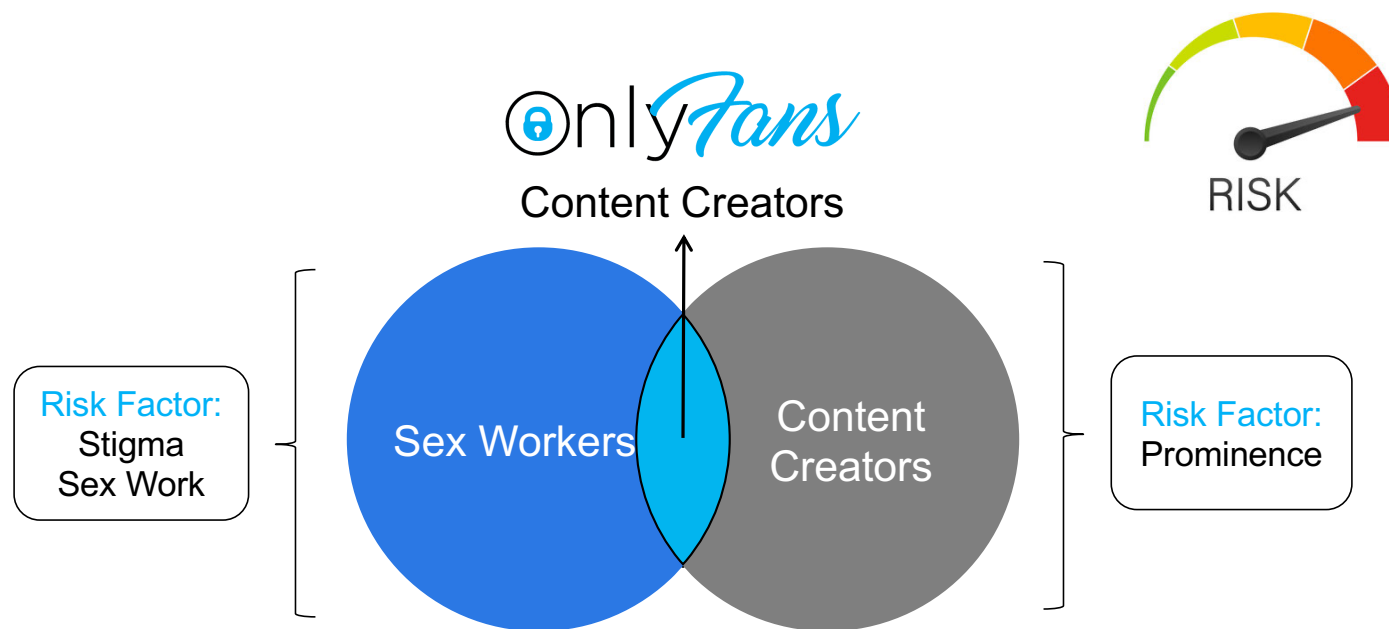
III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

OnlyFans



1291.85% increase
in user base between
2019 and 2021

II. NDII: Intimate Sharing



III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

Another 2-year project: qualitative

“Nudes? Shouldn’t I charge for these?” Motivations of New Sexual Content Creators on OnlyFans. Hamilton, V., Soneji, A., McDonald, A.M., Redmiles, E.M.

ACM CHI 2023 Best Paper Honorable Mention



Risk, Resilience and Reward: Impacts of Shifting to Digital Sex Work. Hamilton, V., Barakat, H.L., Redmiles, E.M. ACM CSCW 2022.

“I feel physically safe but not politically safe”: Understanding the Digital Threats and Safety Practices of OnlyFans Creators. Soneji, A., Hamilton, V., Doupe, A., McDonald, A.M., Redmiles E.M. USENIX Security 2024.

RESEARCH QUESTION 1

What motivates starting professional sexual content creation?

RESEARCH QUESTION 2

What risks persist from in-person sex work & what novel risks emerge?

STRATEGY

What strategies do people use to stay safe?

Another 2-year project: qualitative

Global North US, UK, Germany, Sweden, Spain, France, Canada

58 Interviews with 19 people who were new to the sex industry and 37 people who were in-person sex workers and pivoted online

2 Rounds of Sampling at the onset of COVID-19 and from Sept-October 2021.

Protocol, study materials, and recruitment were reviewed by sex working consultants to ensure the appropriateness and ethics of materials.

III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

Another 2-year
project: qualitative

RESEARCH QUESTION 1

What motivates starting professional sexual content creation?

RESEARCH QUESTION 2

What risks persist from in-person sex work & what novel risks emerge?

STRATEGY

What strategies do people use to stay safe?

III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

RESEARCH QUESTION 1: WHAT MOTIVATES STARTING PROFESSIONAL SEXUAL CONTENT CREATION?

We find **5 motivations** for newcomers **to use OnlyFans**

1

Visibility & Acceptance

Mainstream visibility
and acceptance of
OnlyFans as a result
of celebrity hype,
platform design, and
peer suggestion.

Hamilton, V., Soneji, A., McDonald, A.M., Redmiles, E.M. “Nudes? Shouldn’t I charge for these?”:
Motivations of New Sexual Content Creators on OnlyFans. ACM CHI 2023. Best Paper Honorable Mention

III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

RESEARCH QUESTION 1: WHAT MOTIVATES STARTING PROFESSIONAL SEXUAL CONTENT CREATION?

We find **5 motivations** for newcomers **to use OnlyFans**

1

Visibility & Acceptance

Mainstream visibility and acceptance of OnlyFans as a result of celebrity hype, platform design, and peer suggestion.

Celebrity Hype

Cardi B joined in August 2020 after Beyoncé mentioned the site in a remix in April, which caused, according to a spokesperson for OnlyFans, “a 15 percent spike in traffic” in the subsequent 24 hours

III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

RESEARCH QUESTION 1: WHAT MOTIVATES STARTING PROFESSIONAL SEXUAL CONTENT CREATION?

We find **5 motivations** for newcomers **to use OnlyFans**

1

Visibility & Acceptance

Mainstream visibility and acceptance of OnlyFans as a result of celebrity hype, platform design, and peer suggestion.

Celebrity Hype

“I thought it was cool, I liked the idea of these mostly women... reforming sexuality in their own way and being able to claim, reclaim something”

III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

RESEARCH QUESTION 1: WHAT MOTIVATES STARTING PROFESSIONAL SEXUAL CONTENT CREATION?

We find **5 motivations** for newcomers **to use OnlyFans**

1

Visibility & Acceptance

Mainstream visibility and acceptance of OnlyFans as a result of celebrity hype, platform design, and peer suggestion.

Celebrity Hype

“at first I was hearing that it was lucrative... [and then] I kept hearing about [OnlyFans everywhere]... [I] wanted to see what the hype is about”

III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

RESEARCH QUESTION 1: WHAT MOTIVATES STARTING PROFESSIONAL SEXUAL CONTENT CREATION?

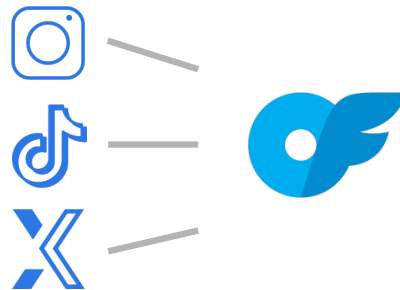
We find **5 motivations** for newcomers **to use OnlyFans**

1

Visibility & Acceptance

Mainstream visibility and acceptance of OnlyFans as a result of celebrity hype, platform design, and peer suggestion.

Platform Design



No discovery forces cross-platform promotion

III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

RESEARCH QUESTION 1: WHAT MOTIVATES STARTING PROFESSIONAL SEXUAL CONTENT CREATION?

We find **5 motivations** for newcomers **to use OnlyFans**

1

Visibility & Acceptance

Mainstream visibility and acceptance of OnlyFans as a result of celebrity hype, platform design, and peer suggestion.

Platform Design

“on every big tweet that blows up, there’s usually someone’s OnlyFans invite.”

III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

RESEARCH QUESTION 1: WHAT MOTIVATES STARTING PROFESSIONAL SEXUAL CONTENT CREATION?

We find **5 motivations** for newcomers **to use OnlyFans**

1

Visibility & Acceptance

Mainstream visibility and acceptance of OnlyFans as a result of celebrity hype, platform design, and peer suggestion.

Peer Suggestion

“Yeah my sister, she’s like you have a great body why don’t you just try OnlyFans? I was only going to do it for maybe just a couple of weeks just to help take care of things financially, but it lasted longer than that.”

III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

RESEARCH QUESTION 1: WHAT MOTIVATES STARTING PROFESSIONAL SEXUAL CONTENT CREATION?

We find **5 motivations** for newcomers **to use OnlyFans**

1

Visibility & Acceptance

Mainstream visibility and acceptance of OnlyFans as a result of celebrity hype, platform design, and peer suggestion.

2

Labor Improvement

Perception of higher, safer, easier platform earnings than other gig and service work.

III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

RESEARCH QUESTION 1: WHAT MOTIVATES STARTING PROFESSIONAL SEXUAL CONTENT CREATION?

We find **5 motivations** for newcomers **to use OnlyFans**

1

Visibility & Acceptance

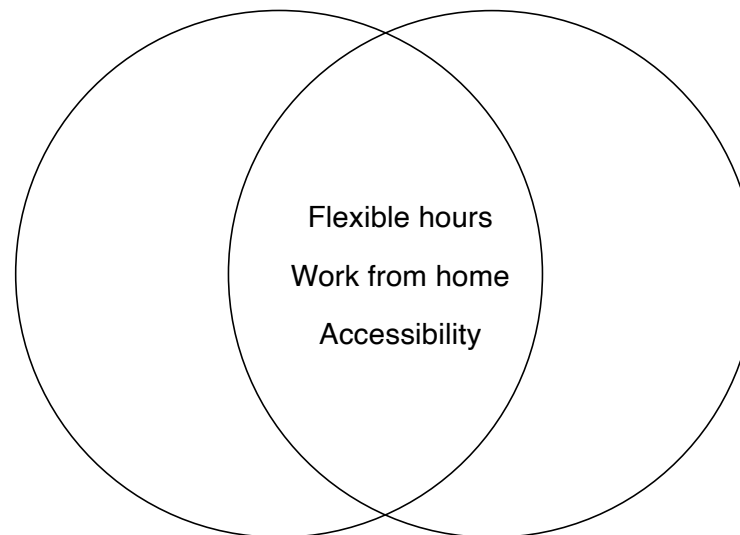
Mainstream visibility and acceptance of OnlyFans as a result of celebrity hype, platform design, and peer suggestion.

2

Labor Improvement

Perception of higher, safer, easier platform earnings than other gig and service work.

Similar to other gig work



III. GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

RESEARCH QUESTION 1: WHAT MOTIVATES STARTING PROFESSIONAL SEXUAL CONTENT CREATION?

We find **5 motivations** for newcomers **to use OnlyFans**

1

Visibility & Acceptance

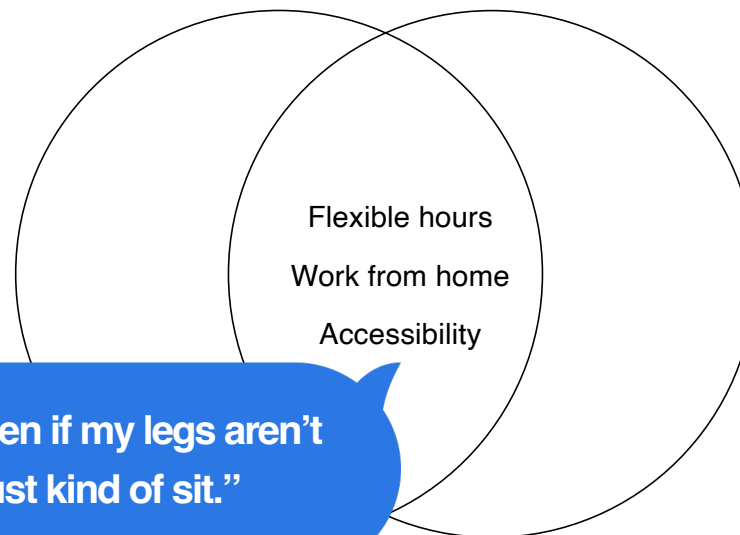
Mainstream visibility and acceptance of OnlyFans as a result of celebrity hype, platform design, and peer suggestion.

2

Labor Improvement

Perception of higher, safer, easier platform

Similar to other gig work



“I can still do OnlyFans even if my legs aren’t working... because I can just kind of sit.”

III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

RESEARCH QUESTION 1: WHAT MOTIVATES STARTING PROFESSIONAL SEXUAL CONTENT CREATION?

We find **5 motivations** for newcomers **to use OnlyFans**

1

Visibility & Acceptance

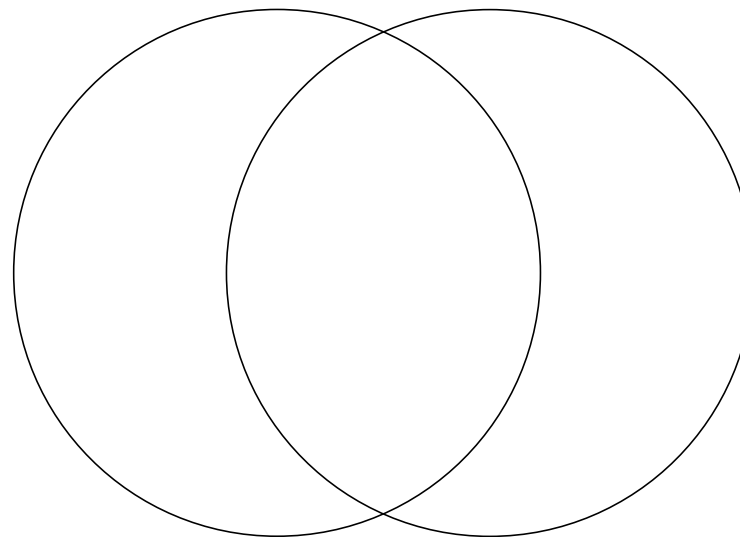
Mainstream visibility and acceptance of OnlyFans as a result of celebrity hype, platform design, and peer suggestion.

2

Labor Improvement

Perception of higher, safer, easier platform earnings than other gig and service work.

Better than other gig work



OnlyFans

III. GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

RESEARCH QUESTION 1: WHAT MOTIVATES STARTING PROFESSIONAL SEXUAL CONTENT CREATION?

We find **5 motivations** for newcomers **to use OnlyFans**

1

Visibility &
Acceptance

2

Labor
Improvement

Better than other gig work

more
control

“allowed for me to establish my boundaries better, unlike the customer service jobs I’ve worked before where I can’t talk back, I can’t cut out a customer because it’s up to the manager, whereas [on OnlyFans] I have the control to just block someone.”

OnlyFans

III. GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

RESEARCH QUESTION 1: WHAT MOTIVATES STARTING PROFESSIONAL SEXUAL CONTENT CREATION?

We find **5 motivations** for newcomers **to use OnlyFans**

1

Visibility & Acceptance

Mainstream visibility and acceptance of

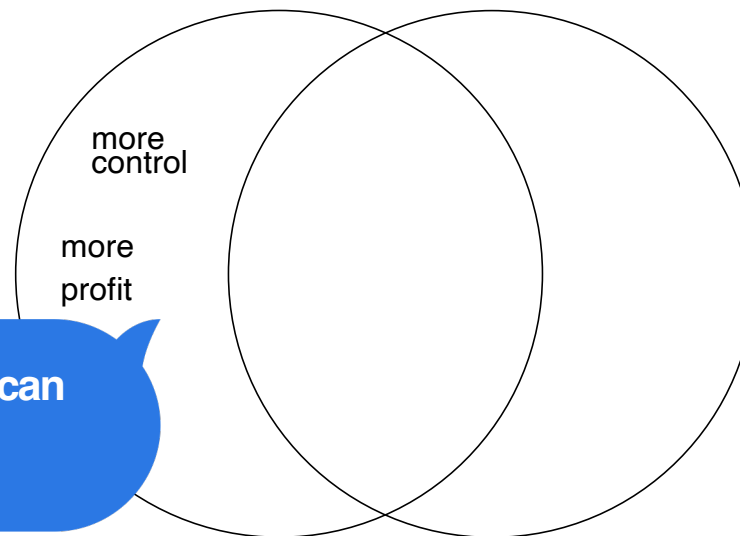
OnlyFans
of celebrities
platform
peer support

2

Labor Improvement

Perception of higher, safer, easier platform

Better than other gig work



“I make more money on OF... And I can control the price more.”

OnlyFans

III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

RESEARCH QUESTION 1: WHAT MOTIVATES STARTING PROFESSIONAL SEXUAL CONTENT CREATION?

We find **5 motivations** for newcomers **to use OnlyFans**

1

Visibility & Acceptance

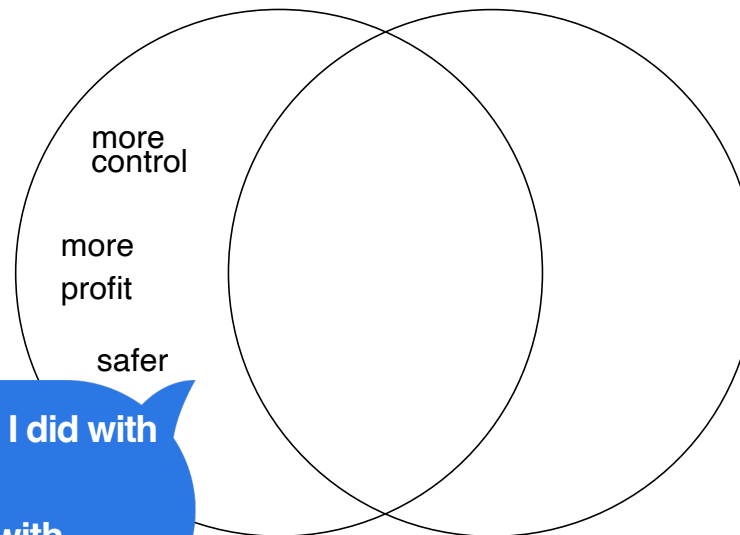
Mainstream visibility and acceptance of OnlyFans as a result of celebrity hype,

2

Labor Improvement

Perception of higher, safer, easier platform earnings than other gig and service work.

Better than other gig work



“I definitely make more money on OnlyFans than I did with Uber and Lyft... and I feel safer because I’m not actually meeting with people. Before, strangers were getting into my car. “

OnlyFans

III. GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

RESEARCH QUESTION 1: WHAT MOTIVATES STARTING PROFESSIONAL SEXUAL CONTENT CREATION?

We find **5 motivations** for newcomers **to use OnlyFans**

1

Visibility & Acceptance

Mainstream visibility and acceptance of OnlyFans as a result of celebrity hype, platform design, and peer suggestion.

2

Labor Improvement

Perception of higher, safer, easier platform earnings than other gig and service work.

3

Pandemic Factors

Such as increased flexible time, safety concerns about other forms of gig work, and loss of other forms of income due to economic impacts.

III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

RESEARCH QUESTION 1: WHAT MOTIVATES STARTING PROFESSIONAL SEXUAL CONTENT CREATION?

We find **5 motivations** for newcomers **to use OnlyFans**

1

Visibility & Acceptance

Mainstream visibility and acceptance of OnlyFans as a result of celebrity hype, platform design, and peer suggestion.

2

Labor Improvement

Perception of higher, safer, easier platform earnings than other gig and service work.

3

Pandemic Factors

Such as increased flexible time, safety concerns about other forms of gig work, and loss of other forms of income due to economic impacts.

4

Profit Potential

Ability to profit from from existing content, audiences, and/or skills.

III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

RESEARCH QUESTION 1: WHAT MOTIVATES STARTING PROFESSIONAL SEXUAL CONTENT CREATION?

We find **5 motivations** for newcomers **to use OnlyFans**

1

Visibility & Acceptance

Mainstream visibility and acceptance of OnlyFans as a result of celebrity hype, platform design, and peer suggestion.

2

Labor Improvement

Perception of higher, safer, easier platform earnings than other gig and service work.

3

Pandemic Factors

Such as increased flexible time, safety concerns about other forms of gig work, and loss of other forms of income due to economic impacts.

4

Profit Potential

Ability to profit from existing content, audiences, and/or skills.

5

Sexual Expression

Freedom to express oneself in a way often censored from digital spaces & to form community with others around sexual expression.

III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

RESEARCH QUESTION 1: WHAT MOTIVATES STARTING PROFESSIONAL SEXUAL CONTENT CREATION?

We find **5 motivations** for newcomers **to use OnlyFans**

1

Visibility & Acceptance

Mainstream visibility and acceptance of OnlyFans as a result of celebrity hype, platform design, and peer suggestion.

2

Labor Improvement

Perception of higher, safer, easier platform earnings than other gig and service work.

3

Pandemic Factors

Such as increased flexible time, safety concerns about other forms of gig work, and loss of other forms of income due to economic impacts.

4

Profit Potential

Ability to profit from existing content, audiences, and/or skills.

5

Sexual Expression

Freedom to express oneself in a way often censored from digital spaces & to form community with others around sexual expression.

III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

Another 2-year project: qualitative

Hamilton, V., Barakat, H.L., Redmiles, E.M. Risk, Resilience and Reward: Impacts of Shifting to Digital Sex Work. ACM CSCW 2022.

Soneji, A., Hamilton, V., Doupe, A., McDonald, A.M., Redmiles E.M. "I feel physically safe but not politically safe": Understanding the Digital Threats and Safety Practices of OnlyFans Creators. To appear in USENIX Security 2024.

RESEARCH QUESTION 1

What motivates starting professional sexual content creation?

RESEARCH QUESTION 2

What risks persist from in-person sex work & what novel risks emerge?

STRATEGY

What strategies do people use to stay safe?

III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

RESEARCH QUESTION 2

WHAT RISKS PERSIST FROM IN-PERSON SEX WORK & WHAT NOVEL RISKS EMERGE?

Digital-only work requires **bigger digital footprint**



More platforms



Platform
Censorship

III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

RESEARCH QUESTION 2

WHAT RISKS PERSIST FROM IN-PERSON SEX WORK & WHAT NOVEL RISKS EMERGE?

Digital-only work requires **bigger digital footprint**



More platforms



Platform
Censorship

“It’s frustrating to be constantly living censored both on Twitter and Instagram. It’s a systematic way of silencing people of color and queer folks so we just give up and stop posting which essentially is what happens. It’s effective, because it’s exactly what happened. I gave up”

III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

RESEARCH QUESTION 2

WHAT RISKS PERSIST FROM IN-PERSON SEX WORK & WHAT NOVEL RISKS EMERGE?

Digital-only work requires **bigger digital footprint**



More platforms



Platform
Censorship

them

NEWS

Blame This Anti-LGBTQ+ Group for OnlyFans' Ban on Porn

Formerly known as Morality in Media, the National Center on Sexual Exploitation was a key player in lobbying to restrict porn on the site.

BY MOLLY SPRAYREGEN

August 24, 2021

III. GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

RESEARCH QUESTION 2

WHAT RISKS PERSIST FROM IN-PERSON SEX WORK & WHAT NOVEL RISKS EMERGE?

Digital-only work requires **bigger digital footprint**



More platforms



Platform
Censorship

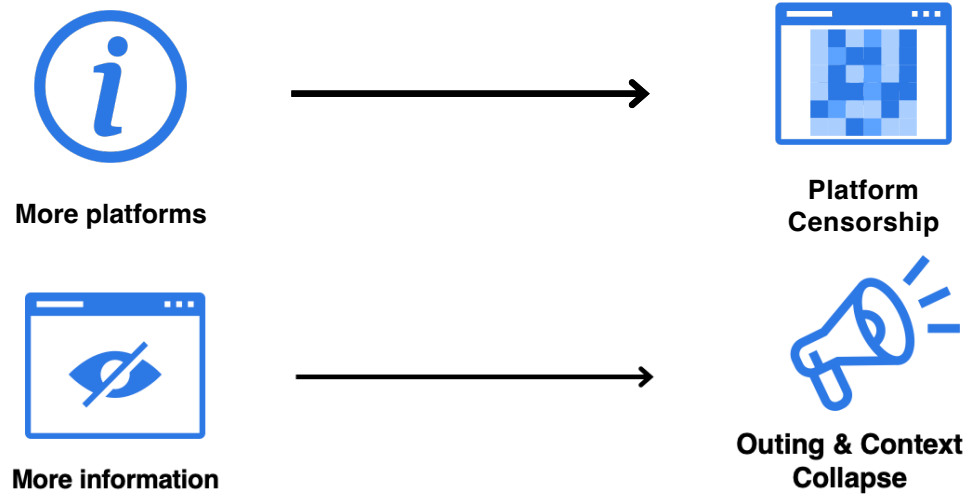
“When you deplatform creators you take away their community and that’s violence. You are isolating people and as we all know from this last year isolation leads to severe mental health consequences”

III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

RESEARCH QUESTION 2

WHAT RISKS PERSIST FROM IN-PERSON SEX WORK & WHAT NOVEL RISKS EMERGE?

Digital-only work requires **bigger digital footprint**

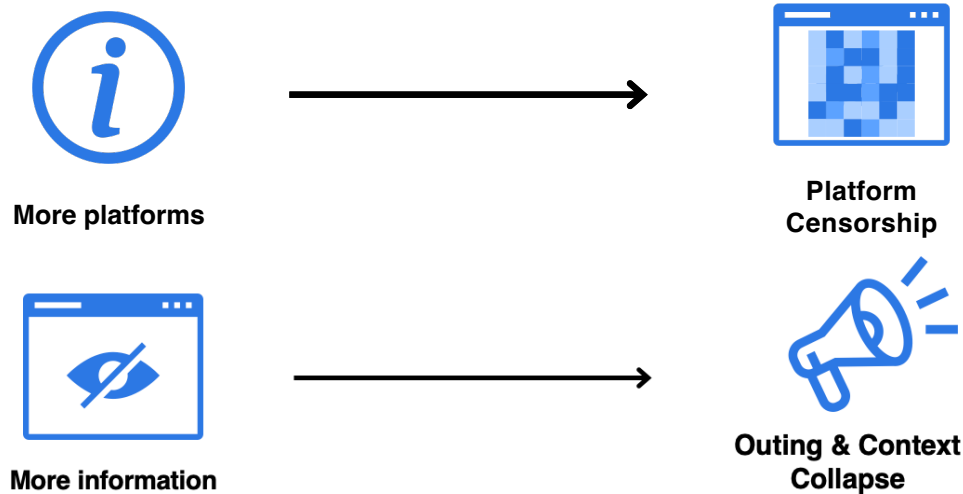


III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

RESEARCH QUESTION 2

WHAT RISKS PERSIST FROM IN-PERSON SEX WORK & WHAT NOVEL RISKS EMERGE?

Digital-only work requires **bigger digital footprint**



“I took a selfie in the bathroom and [uploaded it] to Twitter...a guy went to the same bathroom and posted that selfie of him in that bathroom under mine, like I know where you are... [those] stalker things are the ones that make me more scared.”

Research Participant

III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

RESEARCH QUESTION 2

WHAT RISKS PERSIST FROM IN-PERSON SEX WORK & WHAT NOVEL RISKS EMERGE?

Digital-only work requires **bigger digital footprint**



More platforms



Platform
Censorship



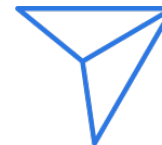
More information



Outing & Context
Collapse

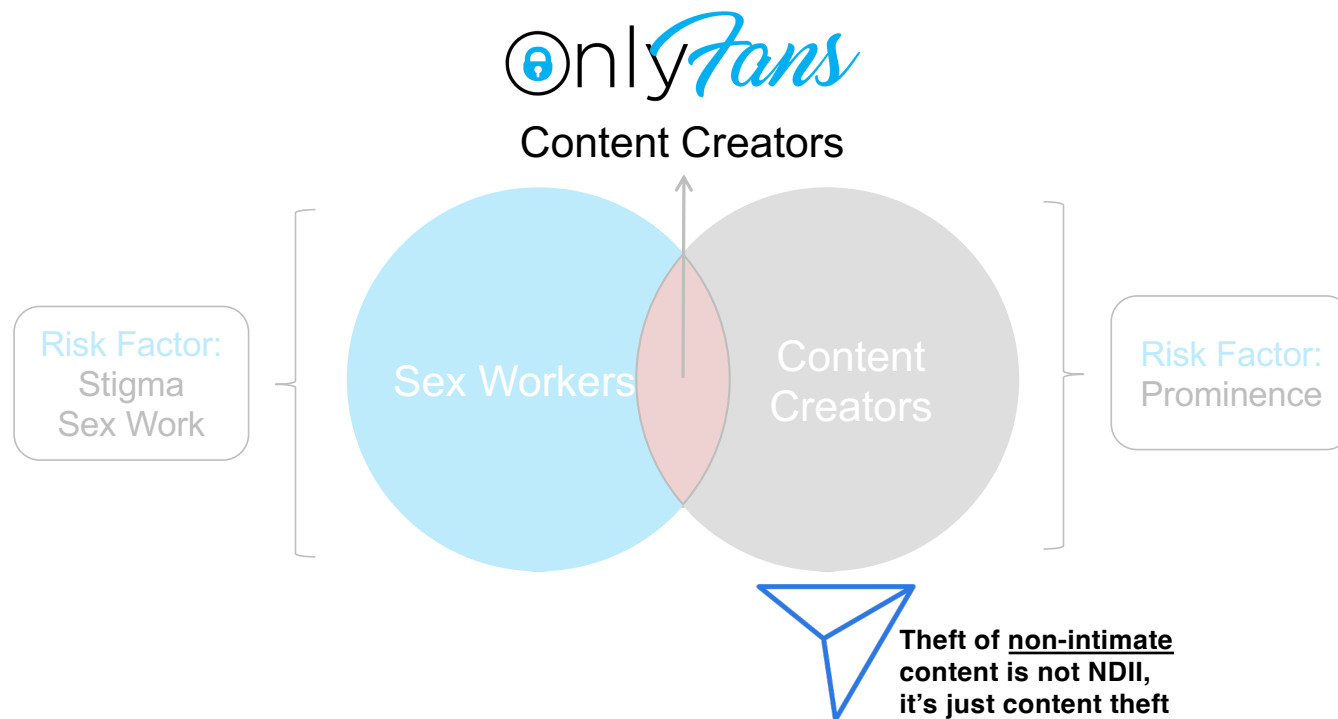


More explicit



Content Theft:
NDII

II. NDII: Intimate Sharing



III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

RESEARCH QUESTION 2

WHAT RISKS PERSIST FROM IN-PERSON SEX WORK & WHAT NOVEL RISKS EMERGE?

Digital-only work requires **bigger digital footprint**

Well, the worst is that you have to have in mind that those videos will be leaked, all [the] people around you are going to get them at some point and they will be forever on the internet...In face-to-face work, I didn't have my intimacy compromised... if I had troubles, they were going to stay there.

Collapse

[With] online work people will...have photos and videos of everything, and I won't be able to ever delete that, they will stay floating around forever.

Also if I stop doing sex work that may affect me in the future too.”

III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

Another 2-year
project: qualitative

RESEARCH QUESTION 1

What motivates starting professional sexual content creation?

RESEARCH QUESTION 2

What risks persist from in-person sex work & what novel risks emerge?

STRATEGY

What strategies do people use to stay safe?

III.GROWTH OF PROFESSIONAL SEXUAL CONTENT CREATION

Another 2-year
project: qualitative

RESEARCH QUESTION 1

What motivates starting professional sexual content creation?

RESEARCH QUESTION 2

What risks persist from in-person sex work & what novel risks emerge?

STRATEGY

What strategies do people use to stay safe?

II. NDII: Intimate Sharing

Defending against recipient resharing

Before Sharing



Rule
Setting



Screening/
Vetting

While Sharing



Expiring
Messages



Screenshot
Notifications



Watermarks

After Sharing



Deletion
Request



Message
Unsend

Defending against identification



Removing
Identifying
Features



Metadata
Removal



Limit Prominence
& Profit



Internet Presence
Monitoring



Separate Online
Identities

IV. OPPORTUNITIES: ENGINEERING & INTERNET GOVERNANCE

What can we build?

IV. OPPORTUNITIES: ENGINEERING & INTERNET GOVERNANCE

What do we need?



Tools that **provide algorithmic transparency**



Tools that **accommodate multiple personas**



Tools that allow for **tracking & protection of content**

IV. OPPORTUNITIES: ENGINEERING & INTERNET GOVERNANCE

Who else can we help?



Tools that provide algorithmic transparency
(e.g., **activists; racial and gender minorities**)



Tools that accommodate multiple personas
(e.g., **LGBTQ users; journalists; activists**)



Tools that allow for tracking & protection of content
(e.g., **Recreational sexters; survivors of intimate image abuse**)

IV. OPPORTUNITIES: ENGINEERING & INTERNET GOVERNANCE

How can we build it?

Risks	Technical Directions
 Platform Censorship	• Filter Analysis • Facial Recognition Circumvention
 Content Leakage	• Certificate Infrastructure for Content • Robust Content Matching • Anti-Theft Technology
 Outing & Context Collapse	• Robust Image Modification • Privacy-Preserving Multi-Profile Support • Automated Blocking / Exposure Measurement • Provable Verification of Privacy Settings

IV. OPPORTUNITIES: ENGINEERING & INTERNET GOVERNANCE

How can we build it?

Risks	Technical Directions
 Platform Censorship	• Filter Analysis • Facial Recognition Circumvention
 Content Leakage	• Certificate Infrastructure for Content • Robust Content Matching • Anti-Theft Technology
 Outing & Context Collapse	• Robust Image Modification • Privacy-Preserving Multi-Profile Support • Automated Blocking / Exposure Measurement • Provable Verification of Privacy Settings



IV. OPPORTUNITIES: ENGINEERING & INTERNET GOVERNANCE

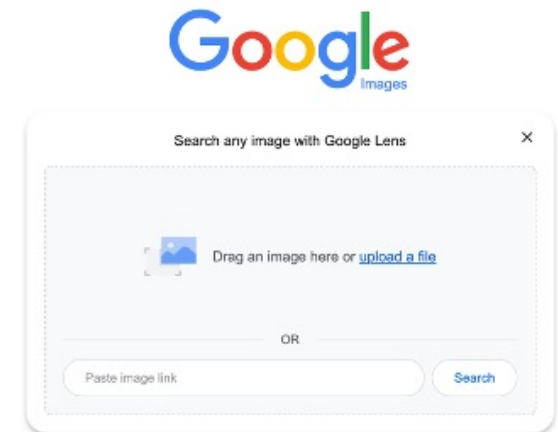
How can we build it?

Risks	Technical Directions
 Platform Censorship	• Filter Analysis • Facial Recognition Circumvention
 Content Leakage	• Certificate Infrastructure for Content • Robust Content Matching • Anti-Theft Technology
 Outing & Context Collapse	• Robust Image Modification • Privacy-Preserving Multi-Profile Support • Automated Blocking / Exposure Measurement • Provable Verification of Privacy Settings

IV. OPPORTUNITIES: ENGINEERING & INTERNET GOVERNANCE

How can we build it?

Risks	Technical Directions
 Platform Censorship	• Filter Analysis • Facial Recognition Circumvention
 Content Leakage	• Certificate Infrastructure for Content • Robust Content Matching • Anti-Theft Technology
 Outing & Context Collapse	• Robust Image Modification • Privacy-Preserving Multi-Profile Support • Automated Blocking / Exposure Measurement • Provable Verification of Privacy Settings



IV. OPPORTUNITIES: ENGINEERING & INTERNET GOVERNANCE

How can we build it?

Risks	Technical Directions
 Platform Censorship	• Filter Analysis • Facial Recognition Circumvention
 Content Leakage	• Certificate Infrastructure for Content • Robust Content Matching • Anti-Theft Technology
 Outing & Context Collapse	• Robust Image Modification • Privacy-Preserving Multi-Profile Support • Automated Blocking / Exposure Measurement • Provable Verification of Privacy Settings

Android Enterprise Help

 Describe your issue

What is an Android Work Profile?

Android 5 or later devices only

An **Android Work Profile** can be set up on an Android device to separate work apps and data from personal apps and data. With a Work Profile you can securely and privately use the same device for work and personal purposes—your organization manages your work apps and data while your personal apps, data, and usage remain private.

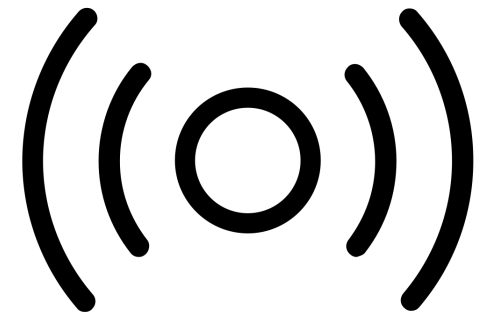
Note: Work Profile apps can't access SMS/MMS data from the personal profile on Android 11+.

If your organization supports enrolling devices to use a Work Profile, your IT department should provide instructions on how to add one to your device.

IV. OPPORTUNITIES: ENGINEERING & INTERNET GOVERNANCE

How can we build it?

Risks	Technical Directions
 Platform Censorship	• Filter Analysis • Facial Recognition Circumvention
 Content Leakage	• Certificate Infrastructure for Content • Robust Content Matching • Anti-Theft Technology
 Outing & Context Collapse	• Robust Image Modification • Privacy-Preserving Multi-Profile Support • Automated Blocking / Exposure Measurement • Provable Verification of Privacy Settings



Some people try to proactively limit their exposure by blocking people from viewing accounts or limiting their prominence.

IV. OPPORTUNITIES: ENGINEERING & INTERNET GOVERNANCE

How can we build it?

Risks	Technical Directions
 Platform Censorship	• Filter Analysis • Facial Recognition Circumvention
 Content Leakage	• Certificate Infrastructure for Content • Robust Content Matching • Anti-Theft Technology
 Outing & Context Collapse	• Robust Image Modification • Privacy-Preserving Multi-Profile Support • Automated Blocking / Exposure Measurement • Provable Verification of Privacy Settings

For more information see:

[Safer Digital Intimacy for Sex Workers and Beyond: A Technical Research Agenda](#)

V. Hamilton, G. Kaptchuk,
A. McDonald and E. M. Redmiles

IEEE Security & Privacy
doi: 10.1109/MSEC.2023.3324615

IV. OPPORTUNITIES: ENGINEERING & INTERNET GOVERNANCE

What can we ~~build~~ do?

IV. OPPORTUNITIES: ENGINEERING & INTERNET GOVERNANCE

1

**Regulate against
globalized
tech censorship of
work & identity**

It has affected sex workers,
trans people, drag queens,
Indigenous communities, etc.

1

Regulate against globalized tech censorship of work & identity

It has affected sex workers,
trans people, drag queens,
Indigenous communities, etc.

“[I want technology] that considers me a person and not a product. But that’s really asking for a lot from the anonymous virtual world.”

IV. OPPORTUNITIES: ENGINEERING & INTERNET GOVERNANCE

1

**Regulate against
globalized
tech censorship of
work & identity**

It has affected sex workers,
trans people, drag queens,
Indigenous communities, etc.

2

**Listen to
marginalized voices
in developing policy**

3

**Empower communities
to design, build &
evaluate the tech that
serves them**

IV. OPPORTUNITIES: ENGINEERING & INTERNET GOVERNANCE

Empower Communities

to design, build &
evaluate the technology
that serves them

“Honestly, I would say one of my biggest pet peeves is that almost all of the platforms that sex workers use as far as advertising and for keeping ourselves safe... none of these are run by sex workers. Many of them are run by older white dudes who are profiting off of the workers. And that I find problematic in many ways”

V. Ideation Time



IV. OPPORTUNITIES: ENGINEERING & INTERNET GOVERNANCE

How can we build it?

Risks	Technical Directions
 Platform Censorship	• Filter Analysis • Facial Recognition Circumvention
 Content Leakage	• Certificate Infrastructure for Content • Robust Content Matching • Anti-Theft Technology
 Outing & Context Collapse	• Robust Image Modification • Privacy-Preserving Multi-Profile Support • Automated Blocking of Contacts
	• Provable Verification of Privacy Settings

II. INTIMATE SHARING PERSONAS

User Personas



Jay



Alice



Peter



Sita



Aliyah

II. INTIMATE SHARING PERSONAS

Remember Jay!

- Jay is a trans woman who shares intimate content with others who are also going through the process of transitioning **to boost body confidence** and mutual support.



she/they

"Early [in] my transition, I had the need to just share some pictures [to] feel more sexy. And so I created an account [on social content platform], that [account] is networked with mostly other trans women that also have these accounts for the exact same purpose."

(Research Participant)

II. INTIMATE SHARING PERSONAS

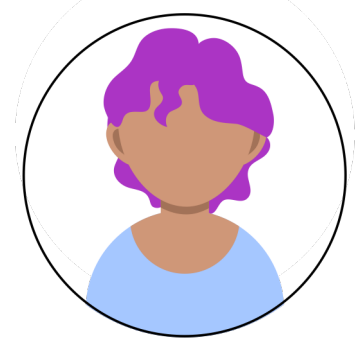
Jay



- Jay is a trans woman who shares intimate content with others who are also going through the process of transitioning to boost body confidence and mutual support.
- Jay is concerned about the potential for violence if members of the religious community she grew up in were to discover her identity as a trans woman. She worries that someone in her body positivity social media group might reshare her intimate content elsewhere.

II. INTIMATE SHARING PERSONAS

Jay



- Jay is a trans woman who shares intimate content with others who are also going through the process of transitioning to boost body confidence and mutual support.
- Jay is concerned about the potential for violence if members of the religious community she grew up in were to discover her identity as a trans woman. She worries that someone in her body positivity social media group might reshare her intimate content elsewhere.
- Jay does not want to be identified through her intimate content. Jay **manually blurs her tattoos and removes metadata** (location, time of image capture, etc.) from her images before posting. She joined a body positivity group that many of her offline friends are in, which has **stated rules** and a moderator.
- **One of Jay's friends recently let her know that she saw her content on Reddit. An NGO is helping Jay get the content off Reddit, but she doesn't know where else it has spread or might re-appear in the future.**

II. INTIMATE SHARING PERSONAS

Remember Peter!

- Peter shares intimate content with potential partners, including strangers, which is a norm on the dating apps he uses for meeting other gay men.



he/him

IV. SOCIOTECHNICAL MITIGATIONS & GAPS

Peter



Peter uses **expiring messages** and **screenshot notifications**, and **unsends a message** after he knows a recipient has seen it.

He recently matched with Rob. After learning they work at the same company, Peter changes his mind about meeting him. Rob begins to send Peter harassing messages.

Peter wants to document these interactions but struggles because Rob uses the same features Peter does. Rob sends expiring messages or unsends a message after Peter views it.

II. INTIMATE SHARING PERSONAS

Peter



- Peter shares intimate content with potential partners, including strangers, which is a norm on the dating apps he uses for meeting other gay men.
- Since he often shares intimate content with strangers, he's worried about one of them re-sharing his images publicly (such as on an online forum) without his consent.

II. INTIMATE SHARING PERSONAS

Peter



“I don’t think there’s ever been a time where I share content and I don’t think about the harm that can be caused.”

(Research Participant)

II. INTIMATE SHARING PERSONAS

Peter



- Peter shares intimate content with potential partners, including strangers, which is a norm on the dating apps he uses for meeting other gay men.
- Since he often shares intimate content with strangers, he's worried about one of them re-sharing his images publicly (such as on an online forum) without his consent.
- He tries to use available features to limit the viewing of his intimate content. He uses **expiring messages**, **screenshot notifications**, and **unsends a message** after he knows a recipient has seen it.
- He recently matched with Rob. After learning they work at the same company, Peter changes his mind about meeting him. Rob begins to send Peter harassing messages. **Peter wants to document these interactions but struggles because of the ephemerality of the content. He could try to screenshot the messages but because of the cryptographic deniability of Signal messages, Rob can just deny that he ever sent the messages.**

II. INTIMATE SHARING PERSONAS

Meet Alice!

- Alice's husband Bob is deployed overseas. They send each other intimate content to stay connected. When Bob is home, they sometimes take photos together during sex to watch while they're apart.



she/her

II. INTIMATE SHARING PERSONAS

Alice



- Alice's husband Bob is deployed overseas. They send each other intimate content to stay connected. When Bob is home, they sometimes take photos together during sex to watch while they're apart.
- Alice and Bob later get divorced and she doesn't want him to have access to her content anymore.

II. INTIMATE SHARING PERSONAS

Alice



“He was vindictive so he likely saved them and showed my pictures to his friends... but even if he didn’t download them I assume they’re still on his phone in our chat history. I just want him to delete them and move on.”

(Research Participant)

II. INTIMATE SHARING PERSONAS

Alice



- Alice's husband Bob is deployed overseas. They send each other intimate content to stay connected. When Bob is home, they sometimes take photos together during sex to watch while they're apart.
- Alice and Bob later get divorced and she doesn't want him to have access to her content anymore.
- Alice sometimes sent photos on WhatsApp, which has a **"Delete for Everyone" option** that she used when her marriage ended.
- **However, she does not know if Bob downloaded them. She also knows that Bob had some photos of her in his own camera roll.**

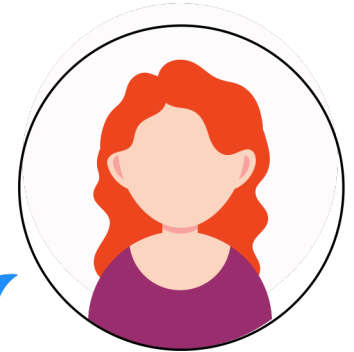
Meet Sita!



she/they

- Sita is an adult content creator on OnlyFans. She has one Instagram she uses to attract OnlyFans subscribers and a personal account she uses to stay in touch with friends and family.
- Sita didn't want her friends and family to know about her work, so she blocked all of them – and everyone they follow -- from her work Instagram.
- One day a subscriber reverse image searched an image on her OnlyFans Instagram account that had graffiti from her small town in the background. He checked her small town's location on Instagram and found her personal Instagram. He started commenting on all her posts. She blocked him, but he retaliated and sent her OnlyFans page link to all her friends and family.

Sita



she/her

“just because I sell my nudes doesn’t mean I don’t have a right to a private life. But I didn’t know what I had to keep secret at the start, you have to learn that as you go along”

(Research Participant)

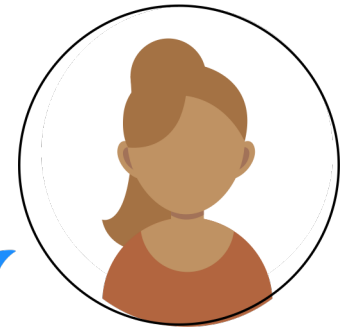
Meet Aliya!



she/her

- Aliya is an in-person sex worker. She dances at a local strip club and occasionally meets a customer from the club for a dinner date or a double date with another worker, at the client's home.
- She uses the most stringent privacy settings on her personal social media and makes sure to always use a VPN when she accesses her sex work social media and email accounts.
- But, her phone still recommends clients from the club as friends on her personal social media apps.

Aliya



she/her

“I just wish I had more control. I need to keep my work and personal lives separate. At the club, if a guy follows you to your car, the security guy will help. If a guy manages to stalk me through my social media and approaches me at the supermarket, there’s no-one to help me there...”

(Research Participant)

IV. Breakouts



Elissa M. Redmiles

 elissa.redmiles@georgetown.edu

 www.elissaredmiles.com



safe.digitalintimacy.org