CONTENT WARNING: harassment, stalking, sexual violence

Digital Safety: A Problem at the Intersection of Security, Engineering, and Society

Veronica A. Rivera ’17

varivera@cs.stanford.edu

https://vrivera2017.github.io/

Stanford University

HAI
Stanford University
Human-Centered
Artificial Intelligence

Stanford University
McCoy Family Center
for Ethics in Society

About me: Veronica Rivera

HARVEY MUDD COLLEGE

BS in CS/Math ’17

UNIVERSITY OF CALIFORNIA SANTA CRUZ

PhD in Computational Media ’23

UF UNIVERSITY of FLORIDA

MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS

Postdoctoral researcher in
CS + Ethics (’23-’25)

Georgia Tech

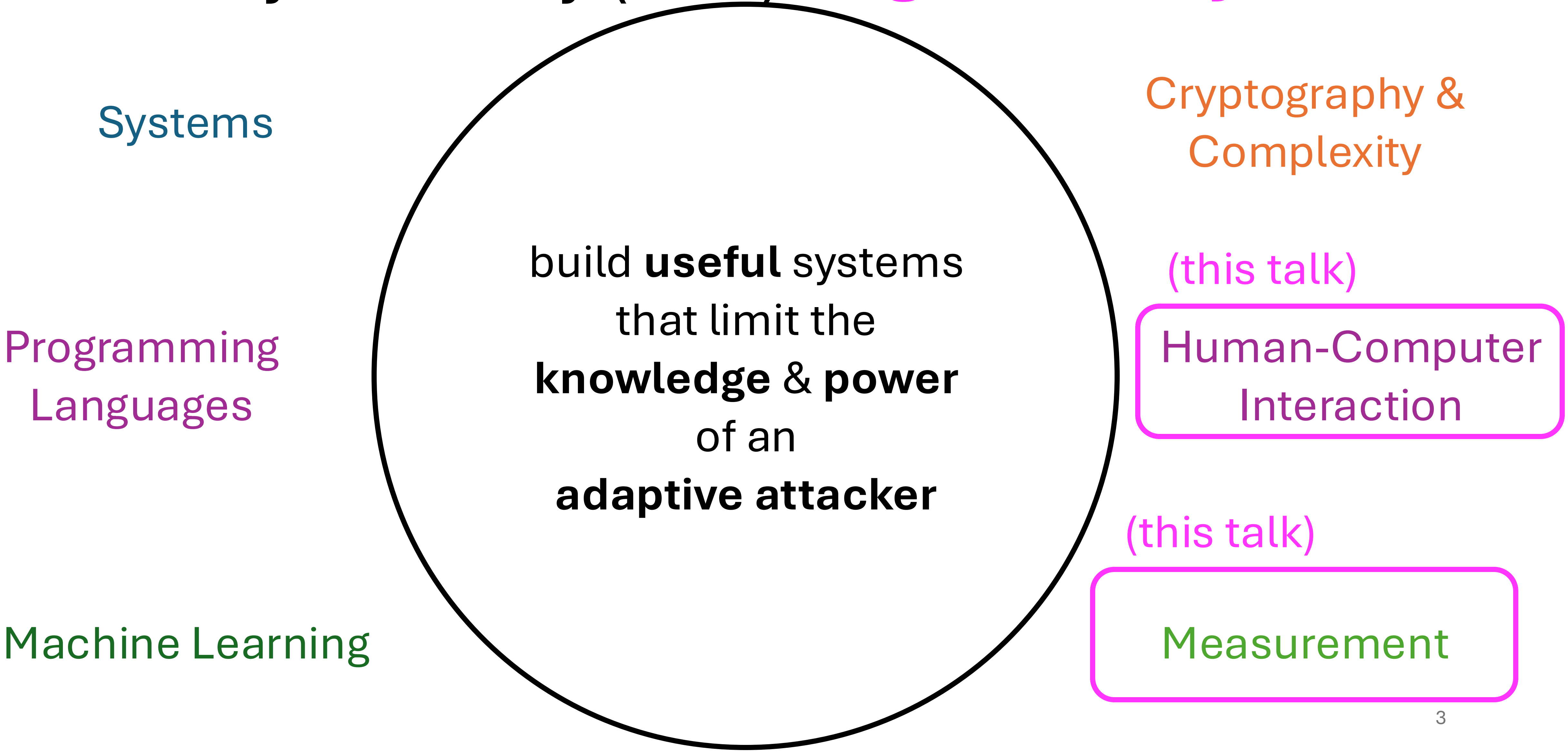
Assistant Professor
(Starting Aug. ’26)

MAX PLANCK INSTITUTE FOR SECURITY AND PRIVACY

Postdoctoral researcher

Fun fact: I did summer research at HMC with Prof. Dodds
in 2015! This got me excited about grad school ☺
I’m also a Claremont native!

Security & Privacy (S&P) :Digital Safety



Systems

Programming Languages

Machine Learning

build **useful** systems that limit the **knowledge** & **power** of an **adaptive attacker**

Cryptography & Complexity

(this talk)

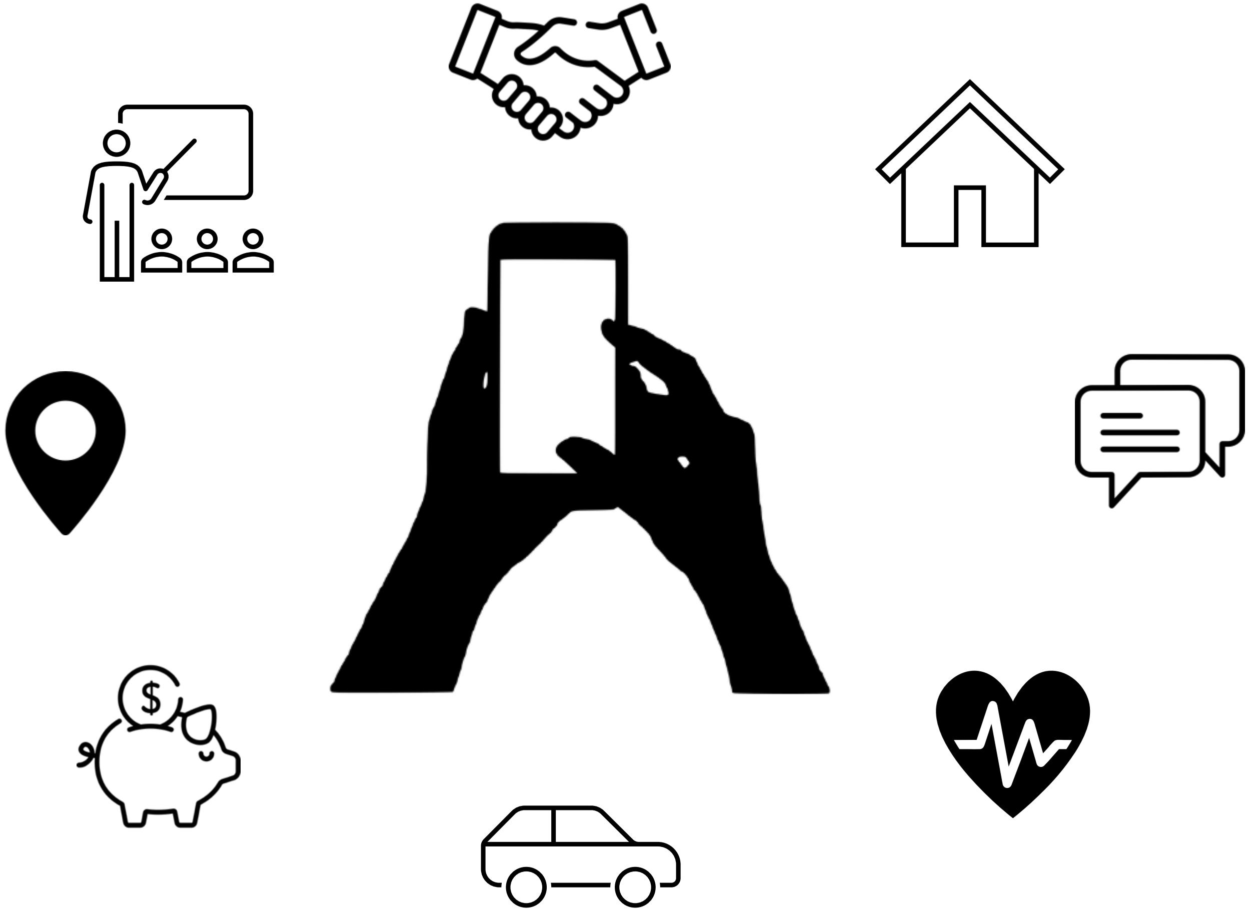
Human-Computer Interaction

(this talk)

Measurement

Sociotechnical systems bridge our digital and physical worlds

A technical system that influences societal dynamics



Sociotechnical systems facilitate harm

Negative impact to people’s physical, psychological, economic, and social well-being, caused by technology

80% of offline stalking is mediated by technology

US Dept. of Justice, 2022

47% of teens who experience online harassment also experience **offline harassment**

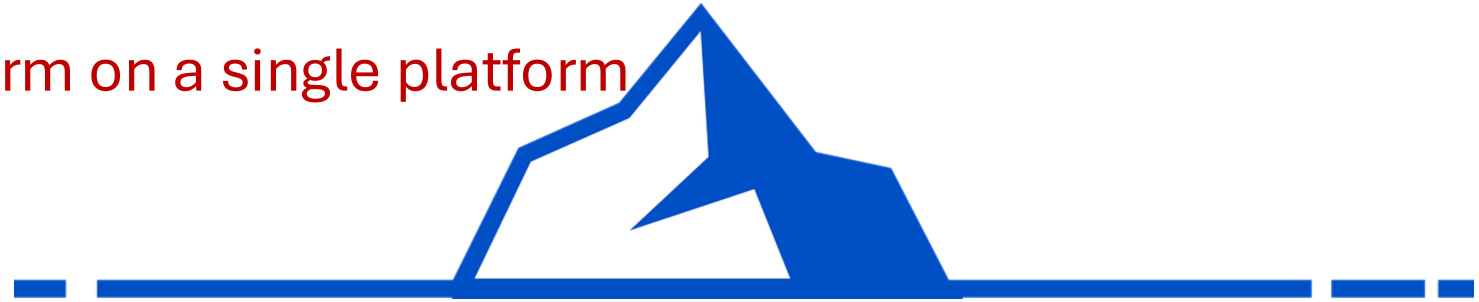
ADL, 2023

75% of dating app users reported experiencing sexual violence

Australian Institute of Criminology, 2024

Existing technical mitigations fall short

Current: Single platform protections

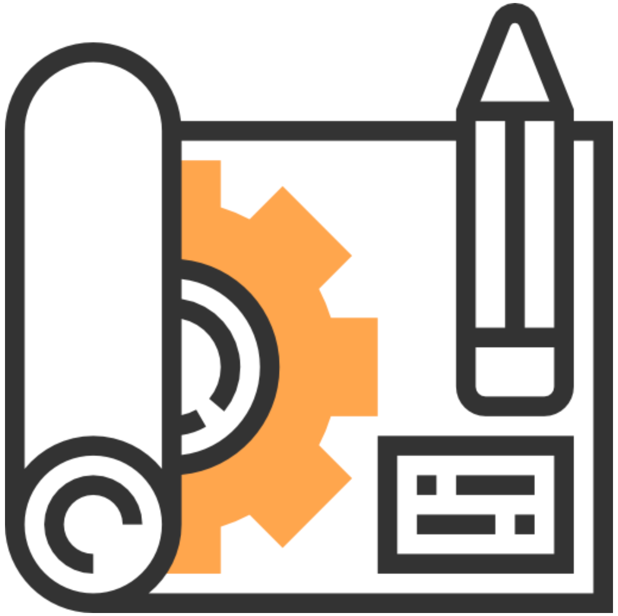


Future: Ecosystem-level protections

We need a blueprint of *cross-platform* harm

Challenges:

- Cross-platform harms are difficult to measure and model
- So we have incomplete data and visibility into these harms

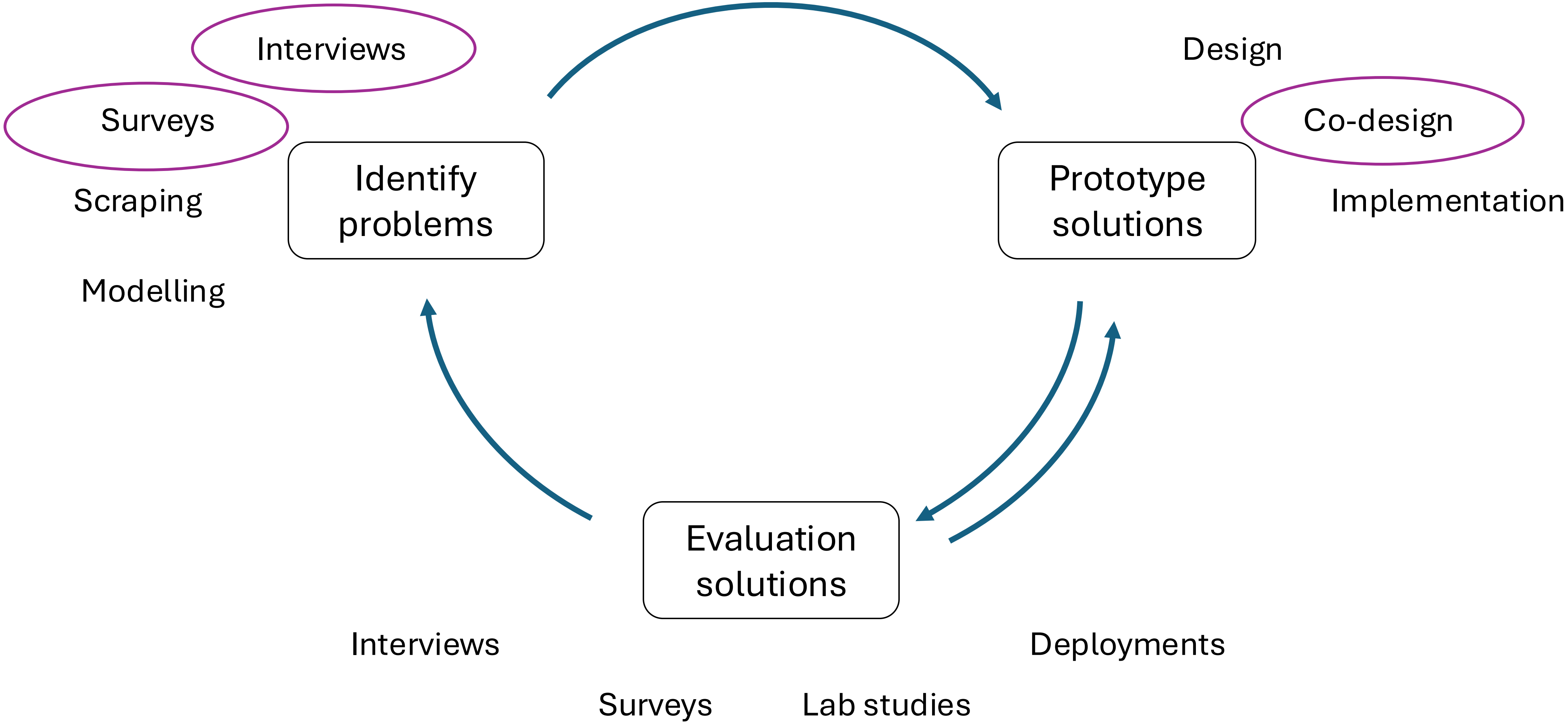


We can build that blueprint from users’ expertise

- How do users self-protect?
- What harm are they’re protecting against?
- How does technology cause harm?

By designing technical protections that support how people are already protecting themselves, we can make sociotechnical systems safer

Empirical & design techniques in engineering



Interviews

Surveys

Scraping

Modelling

Identify problems

Design

Co-design

Implementation

Prototype solutions

Evaluation solutions

Interviews

Surveys

Lab studies

Deployments

Today’s talk:

A framework on/offline harms & protection

Safer abuse reporting systems

Community engagement & education

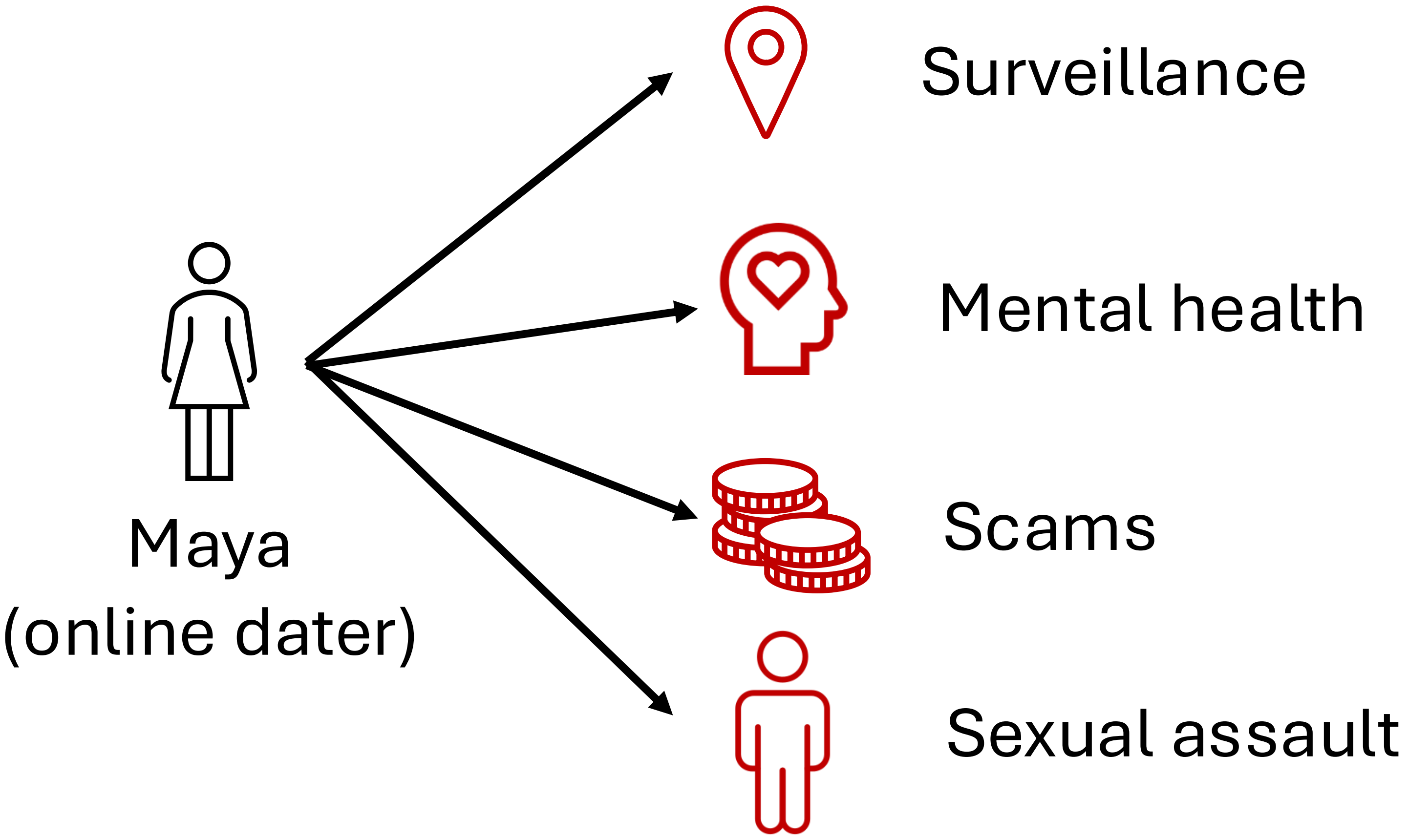
Today's talk:

A framework on/offline harms & protection

Safer abuse reporting systems

Community engagement & education

Online dating platforms facilitate harms



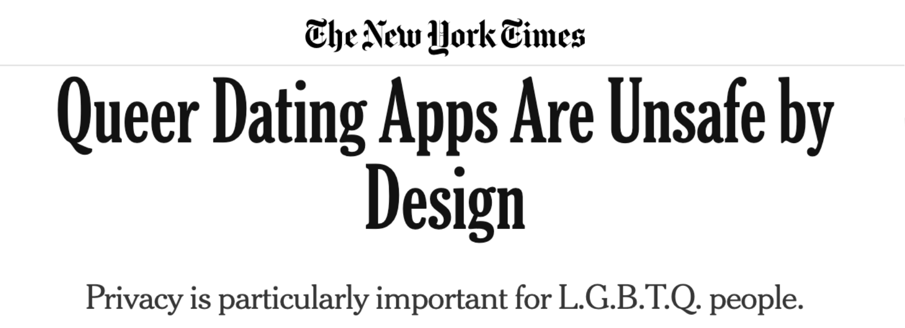
Maya
(online dater)

Surveillance

Mental health

Scams

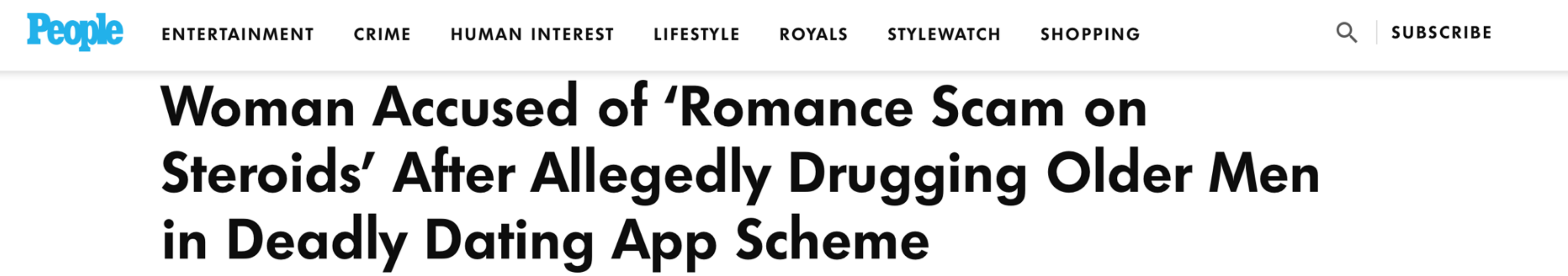
Sexual assault



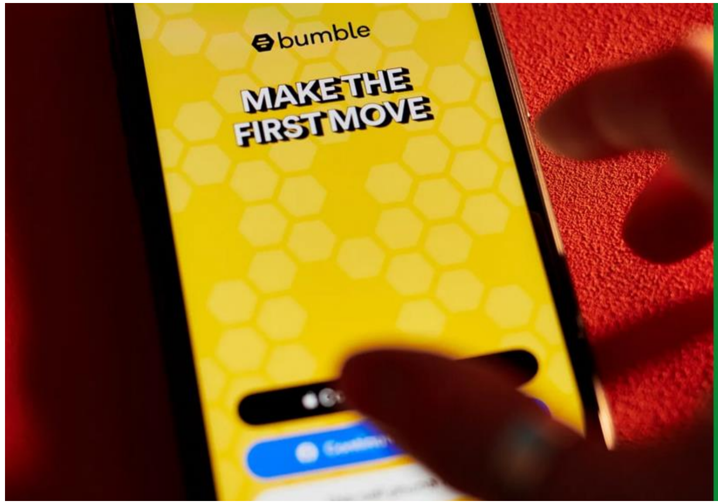
The New York Times

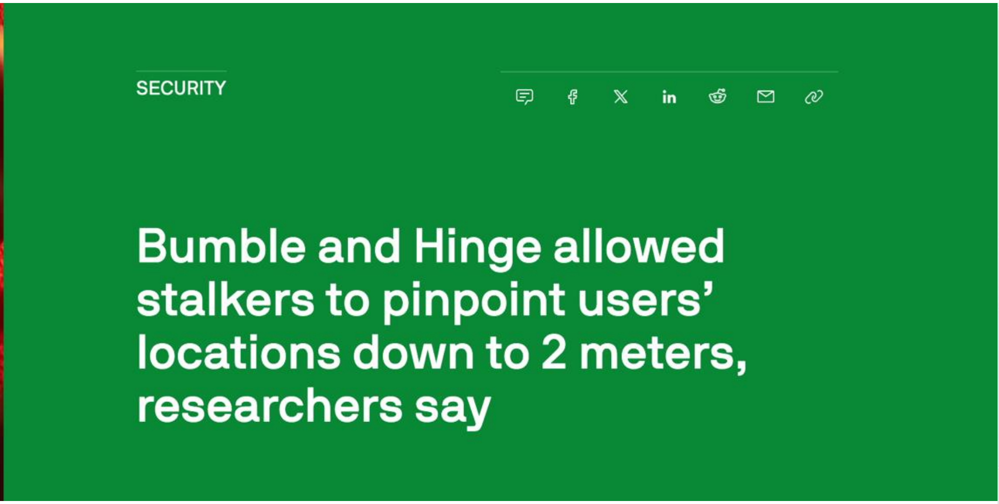
Queer Dating Apps Are Unsafe by Design

Privacy is particularly important for L.G.B.T.Q. people.



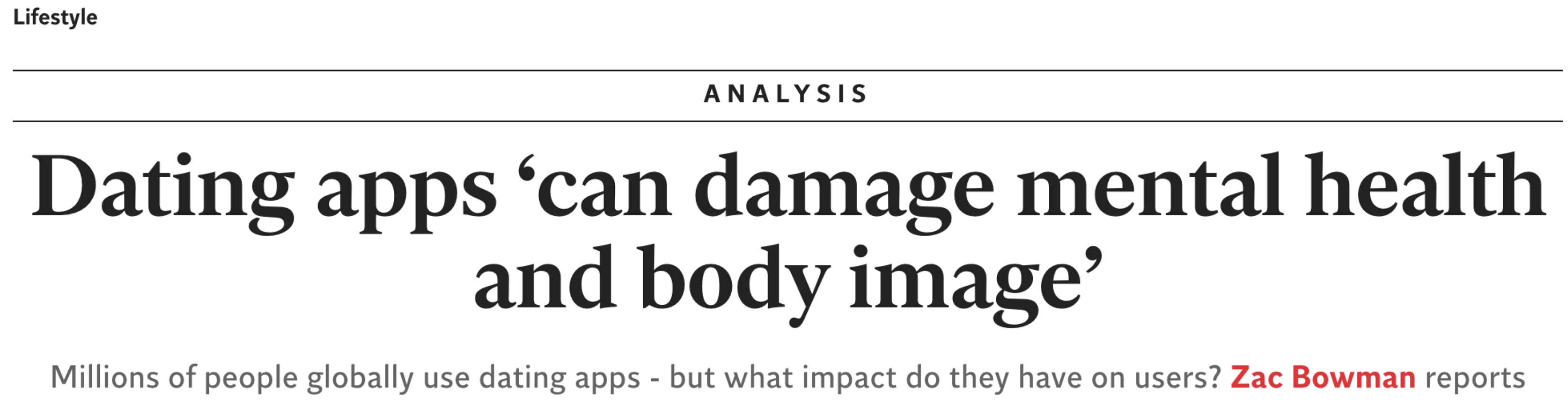
Woman Accused of 'Romance Scam on Steroids' After Allegedly Drugging Older Men in Deadly Dating App Scheme





SECURITY

Bumble and Hinge allowed stalkers to pinpoint users' locations down to 2 meters, researchers say

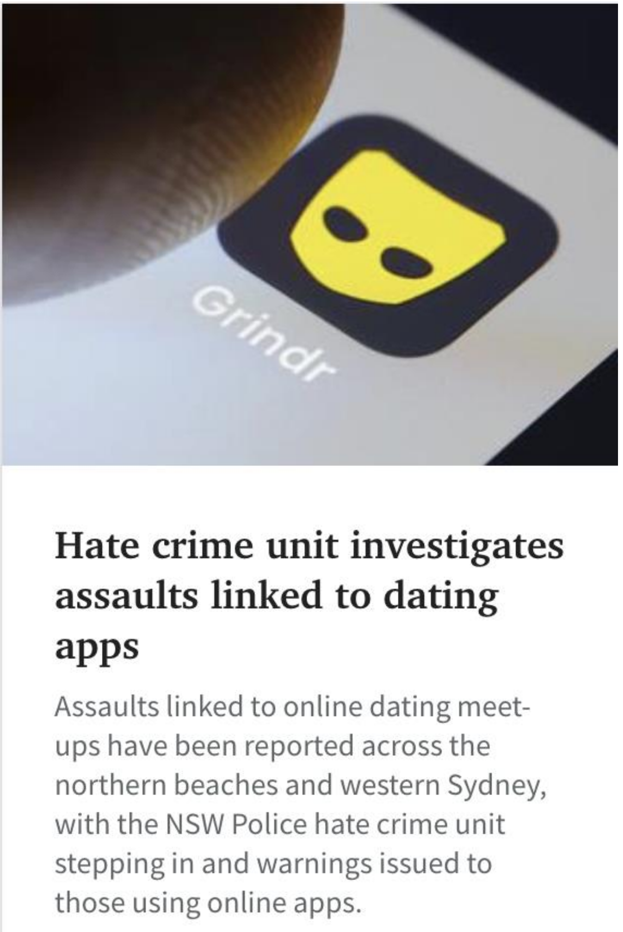


Lifestyle

ANALYSIS

Dating apps 'can damage mental health and body image'

Millions of people globally use dating apps - but what impact do they have on users? **Zac Bowman** reports



Hate crime unit investigates assaults linked to dating apps

Assaults linked to online dating meet-ups have been reported across the northern beaches and western Sydney, with the NSW Police hate crime unit stepping in and warnings issued to those using online apps.

Other groups experience similar digital harms

Journalists

Content creators

Activists

Sex workers

Gig workers

And more…

Granular solutions do not scale

Mitigation for specific populations

Generalized approaches

Goal:
How can we develop a more general understanding of digital safety across user groups and research areas?

Introducing an abstraction across multiple groups

Our solution: Study an abstraction of user groups who leverage similar technologies to accomplish similar goals

Online daters

Sex workers

Gig workers

Veronica A. Rivera, Daricia Wilkinson, Aurelia Augusta, Sophie Li, Elissa M. Redmiles, Angelika Strohmayer. *Safer Algorithmically-Mediated Offline Introductions: Harms and Protective Behaviors*. Published at **CSCW 2024**.

Our abstraction

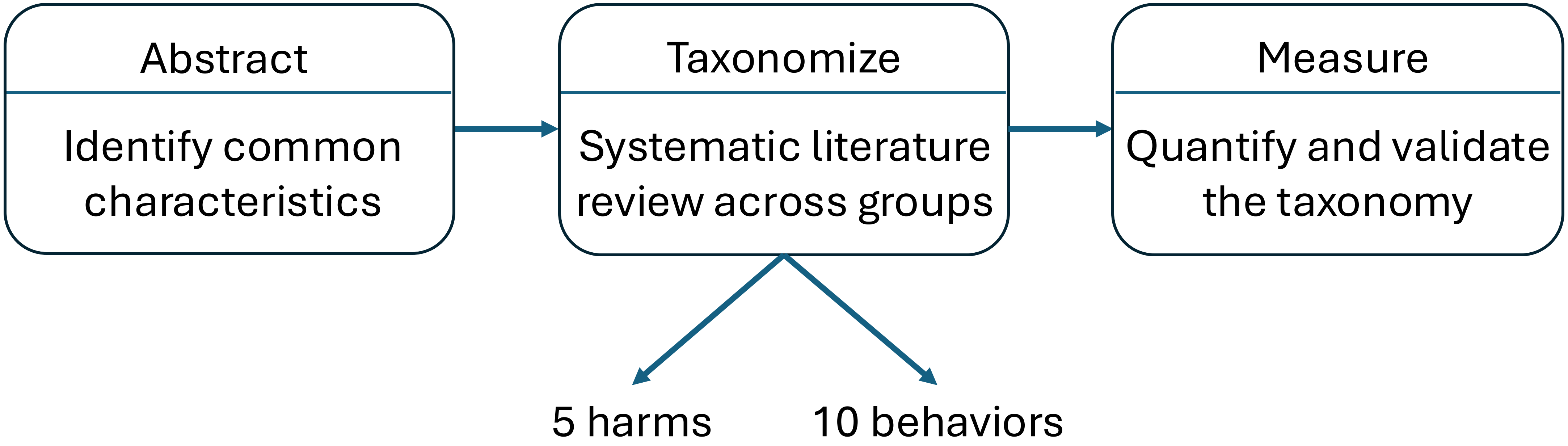
Algorithm matches strangers online

Strangers interact in the physical world

Algorithmically-mediated offline introduction (AMOI):

an **offline** introduction between strangers that is mediated by an **online** matching algorithm on a *digital platform*

Our approach: systematization + measurement



Abstract

Identify common characteristics

Taxonomize

Systematic literature review across groups

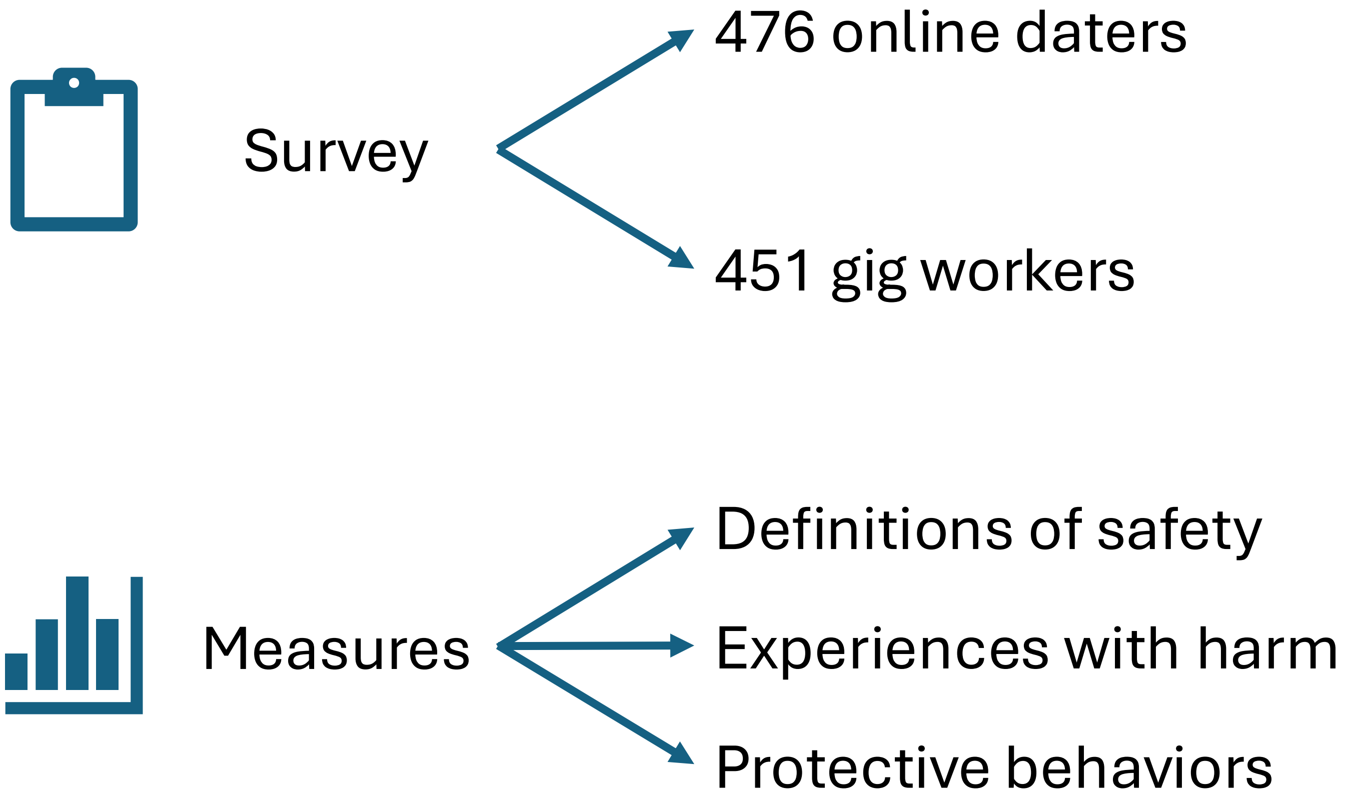
Measure

Quantify and validate the taxonomy

5 harms

10 behaviors

Measure: quantify and validate taxonomy



Survey

476 online daters

451 gig workers

Measures

Definitions of safety

Experiences with harm

Protective behaviors

Five harms

Physical

Financial

Privacy

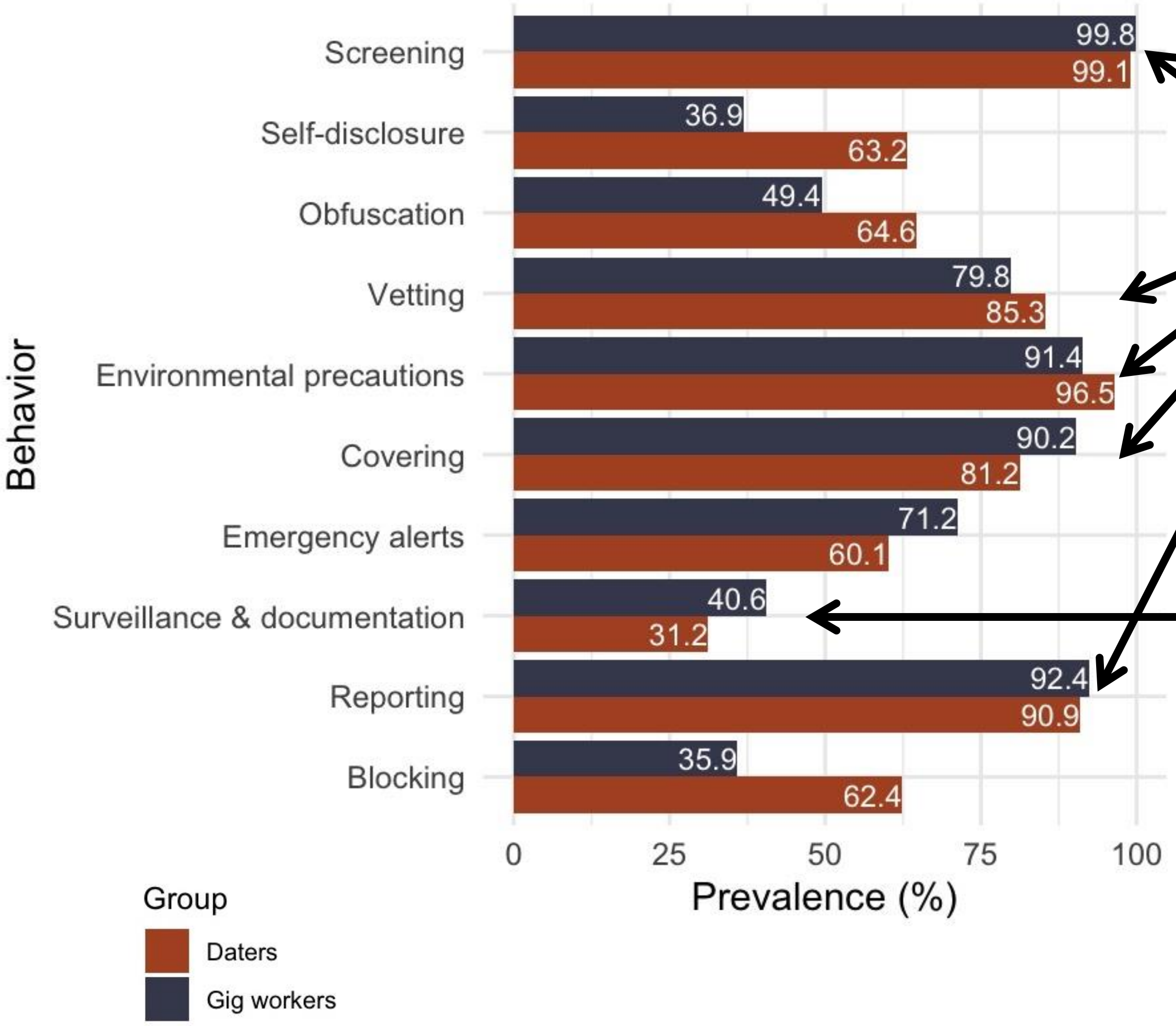
Loss of control over private information

Autonomy

Loss of control over decision-making or physical body

Emotional

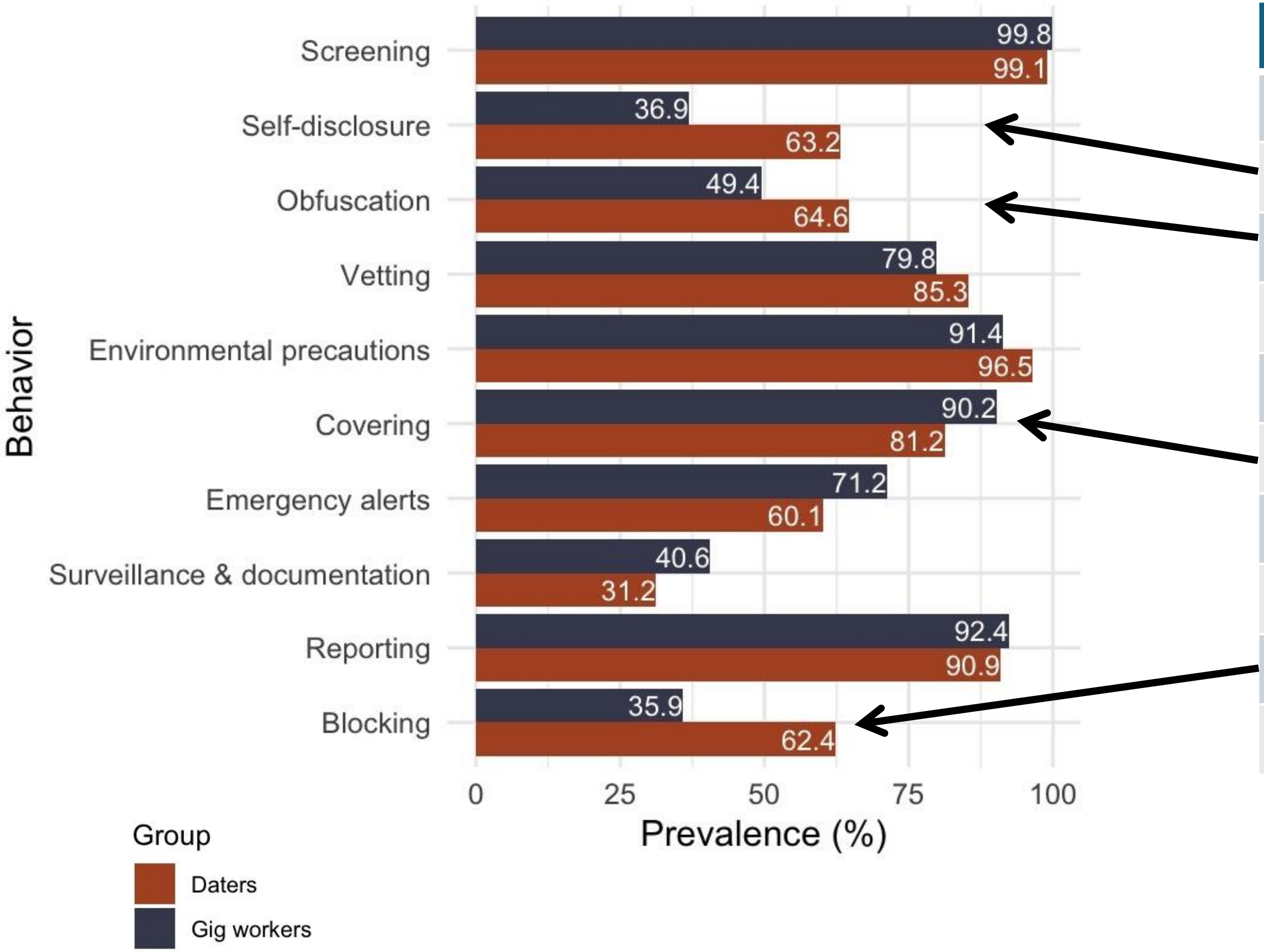
Users self-protection is ***pervasive***



Many behaviors are used by almost all users

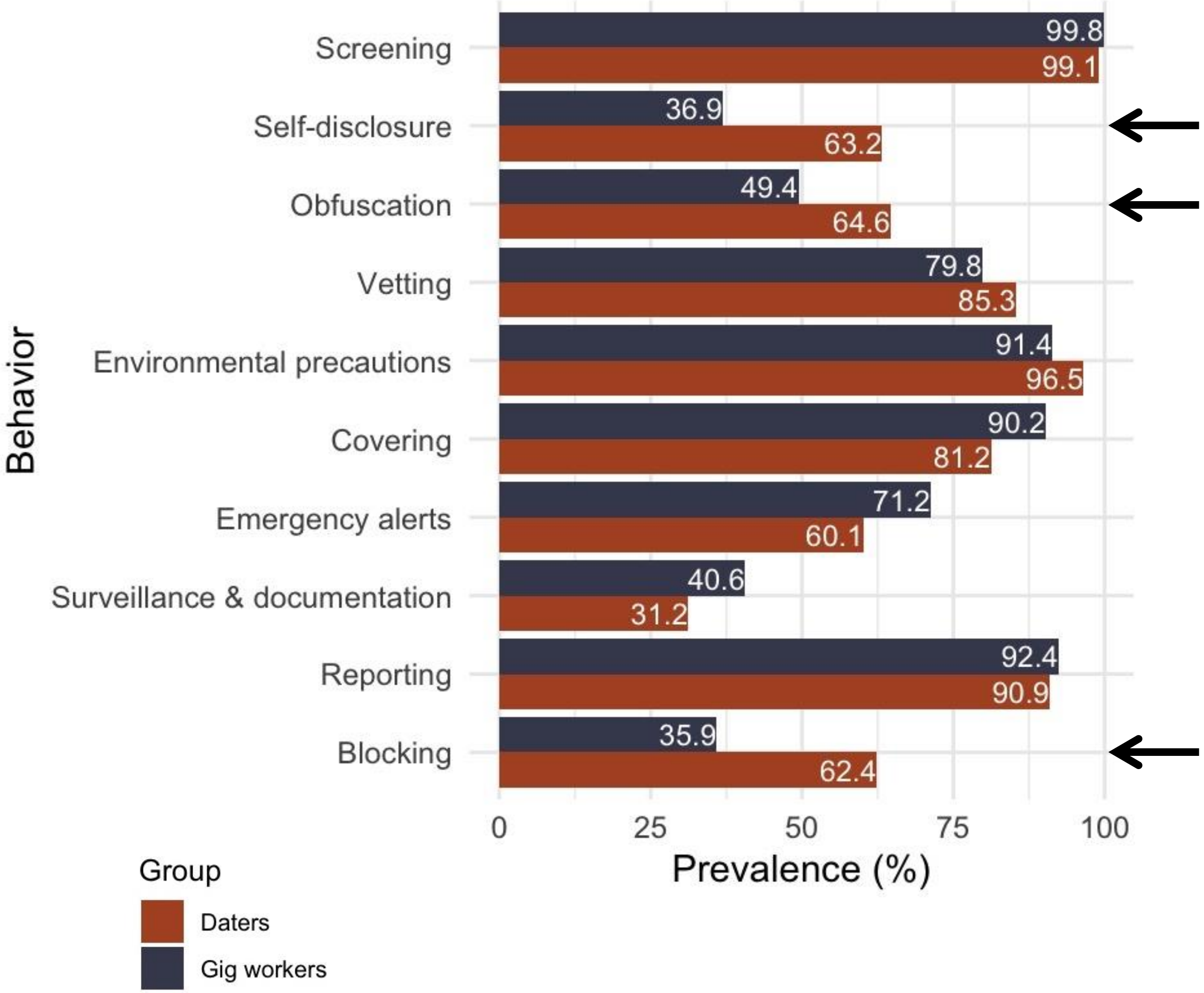
Every behavior is common

Differences are due platform design choices



Behavior	**p-value**
Screening	1.00
Self-disclosure	**< 0.01**
Obfuscation	**< 0.01**
Vetting	1.00
Environmental precautions	0.08
Covering	**0.03**
Emergency alerts	0.07
Documentation	0.10
Blocking	**< 0.01**
Reporting	1.00

Group **differences** suggest design directions



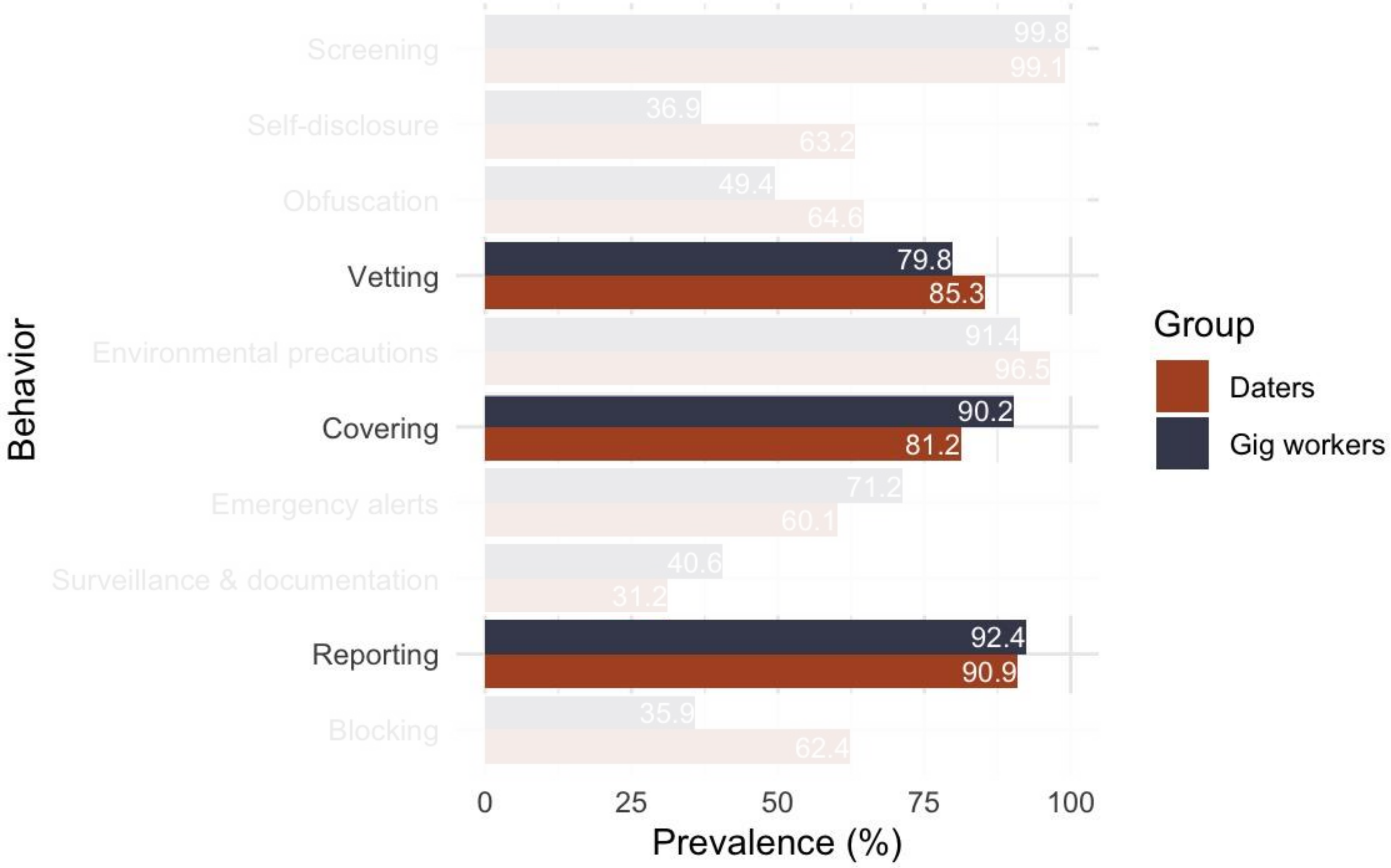
Blocking, self-disclosure, and obfuscation have the biggest **effect size**

Observation: gig platforms don't support these

Group similarities suggest design directions too

≥4/5 respondents use **social** protections:

- **vetting**
- **covering**
- **reporting**



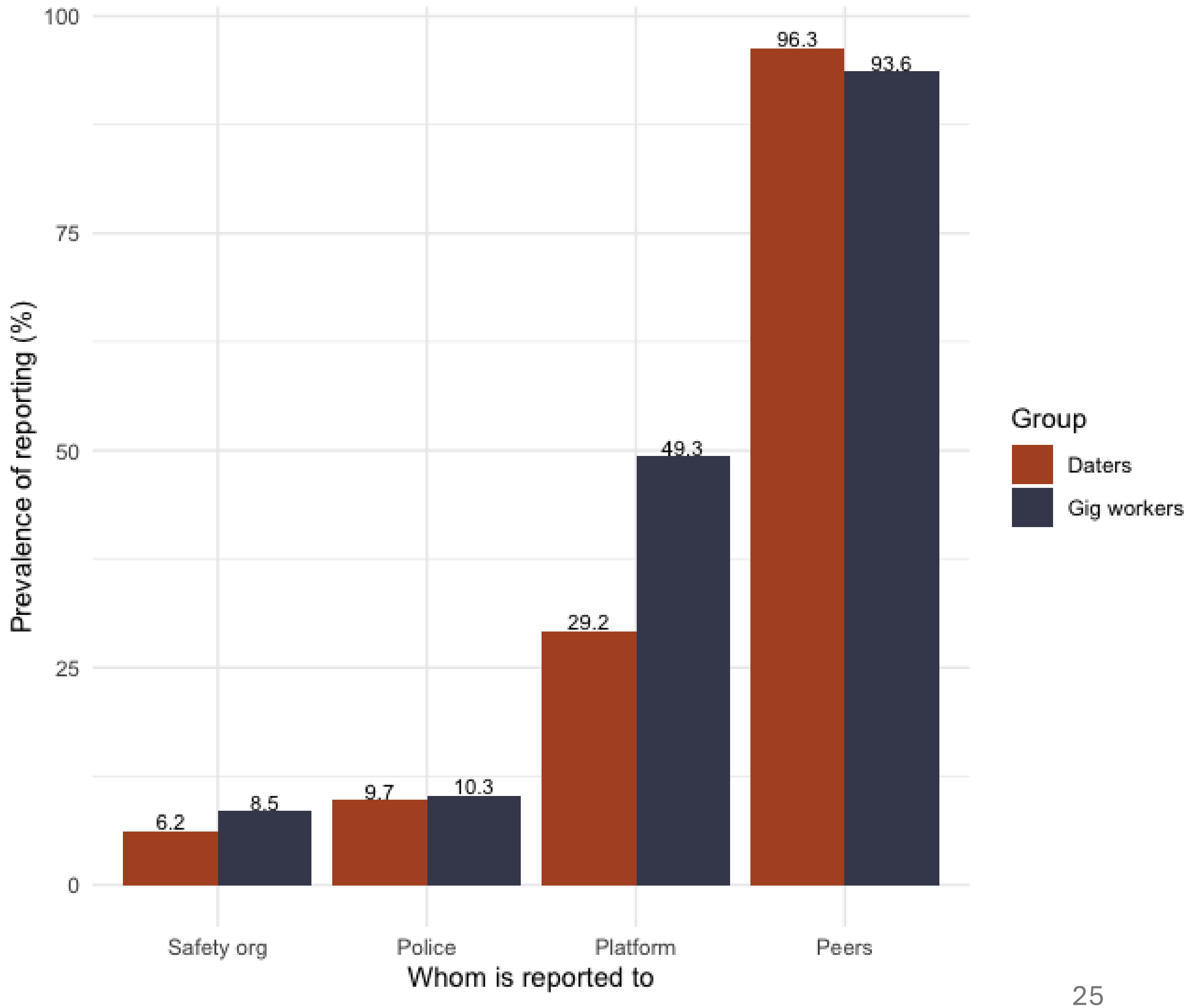
In practice, reporting harm is a **social** behavior

Many report to:

- Peers (>90%)

Few users report to:

- Platforms (< 30%; 50%)
- Police (< 10%)
- Safety NGOs (< 10%)



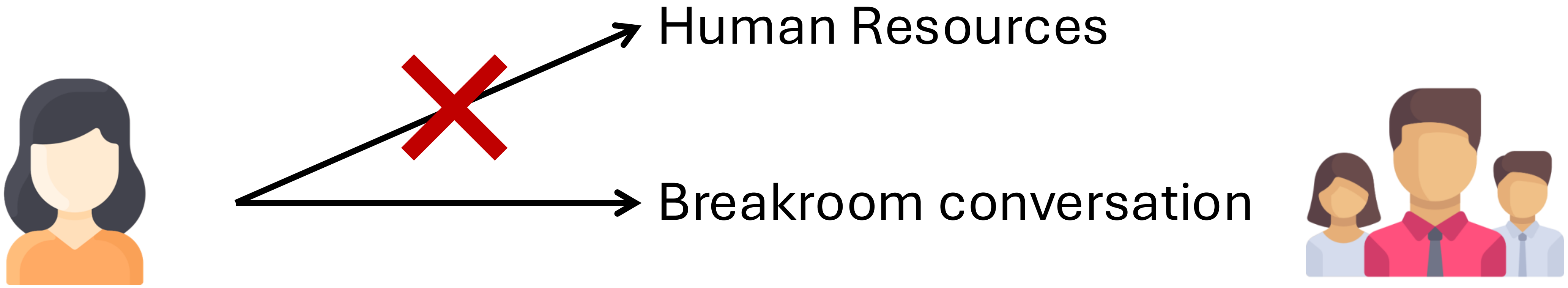
Today’s talk:

A framework on/offline harms & protection

Safer abuse reporting systems

Community engagement & education

Prior work: workplace harassment



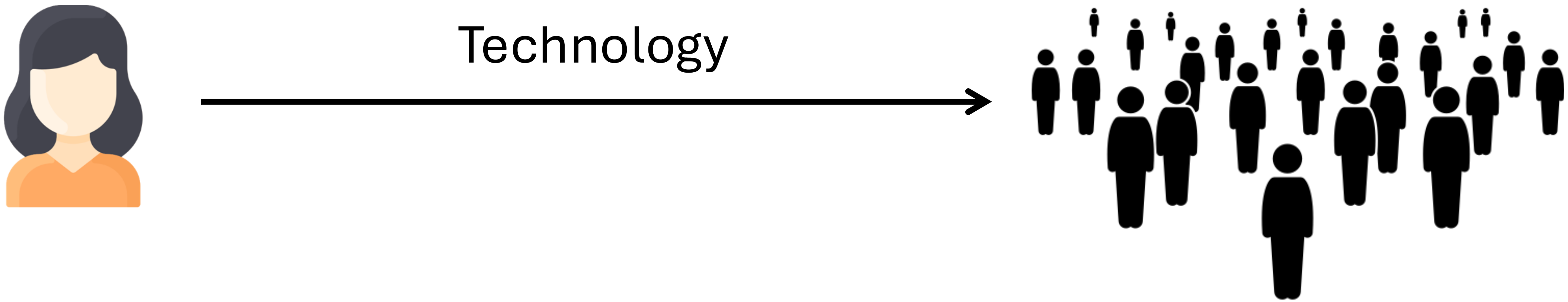
Human Resources

Breakroom conversation

**whisper network: an informal chain of information passed privately between people*

*Johnson, Carrie Ann. "The purpose of whisper networks: a new lens for studying informal communication channels in organizations." *Frontiers in Communication* 8 (2023): 1089335

Digital whisper networks



How do digital whisper networks work?

Research Questions

Goals: What are survivors' goals for reporting experiences with labor abuse to digital whisper networks?

Threats: What are survivors' perceived threats of reporting via digital whisper networks?

Design: How do the goals and threats connect to technological design?

Veronica A. Rivera, Catherine Han, Tracy Li, Elissa M. Redmiles, Zakir Durumeric. *Digital Whisper Networks: Objectives and Threats of Informal Abuse Reporting*. In prep.

Semi-structured interviews

Entertainment

Law

Healthcare

Journalism

Gig work

Technology

Hospitality

Academia

Goals for participating in digital whisper networks

Solicit support
(100%)

Broadcast experience
(88%)

Organize community
(82%)

Passive learning
(76%)

Threats to sharing in digital whisper networks

Uncontrolled re-sharing
(53%)

Retaliation
(41%)

Reputational damage
(35%)

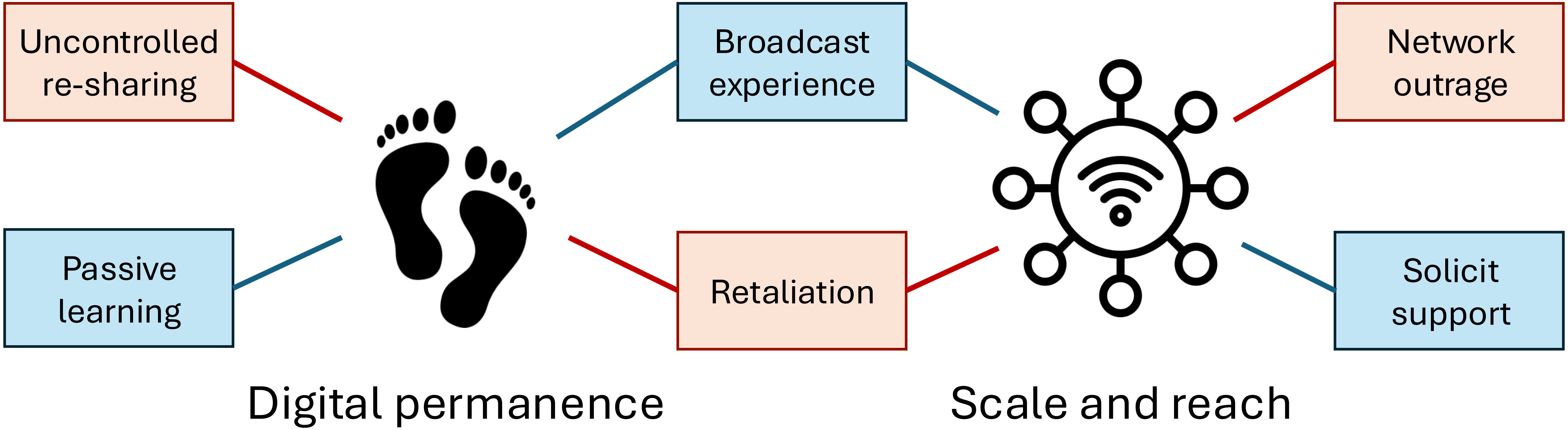
Network damage
(53%)

Access control
(29%)

“They might take the whole text, screenshots, and everything, and send them back to whoever you had a disagreement with.” (P10)

So I guess there there is, of course a concern about “**what if this person who's actually the bad person ends up getting into the group**?” (P15)

Key technical features are contradictory



System design: Blending HCI & Theory

- How can we make peer-to-peer reporting safer?

- Idea: Use cryptographic tools such as **deniable encryption** and **secure reputation systems**

- Research question: Are these tools effective in practice?

Today's talk:

A framework on/offline harms & protection

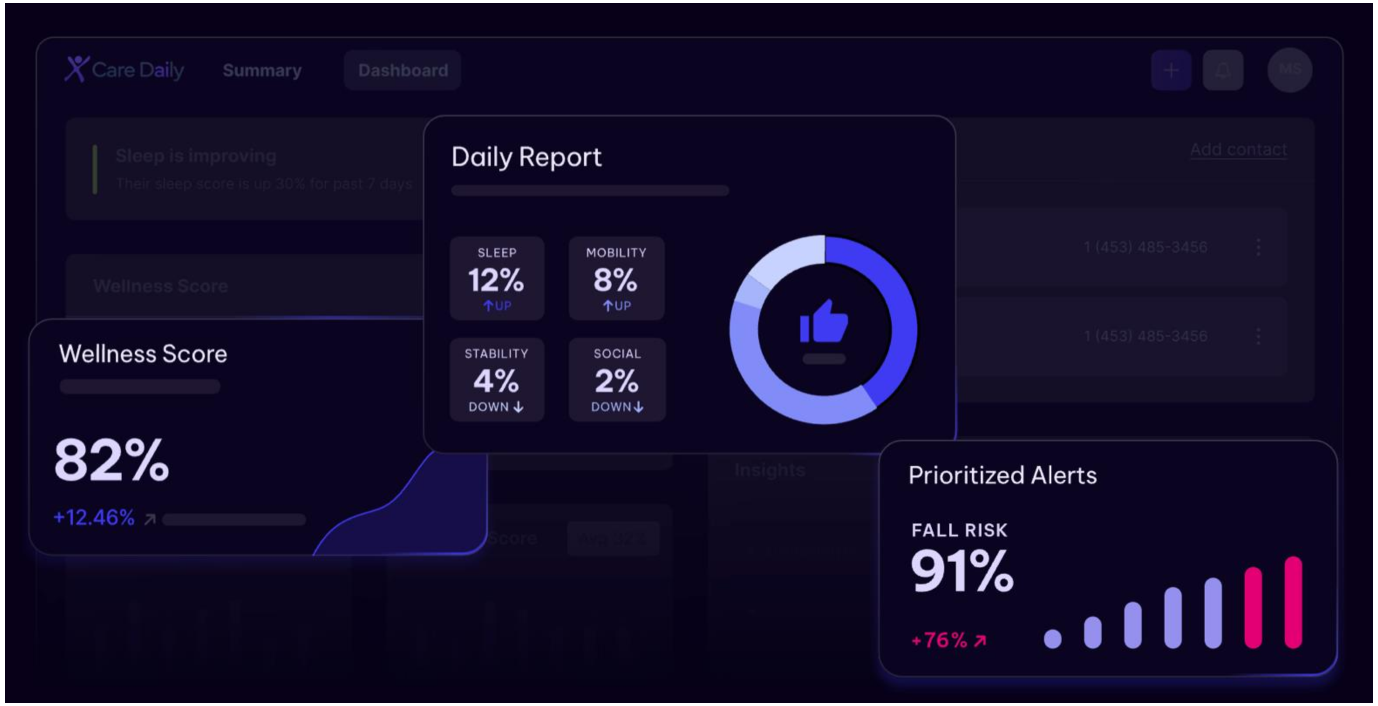
Safer abuse reporting systems

Community engagement & education

AI is shaping labor:

papa
airbnb
TaskRabbit
upwork
DOORDASH
Uber

AI is shaping labor



Sleep is improving

Their sleep score is up 30% for past 7 days

Domestic care work

Source: https://www.domesticworkers.org/about-domestic-work/

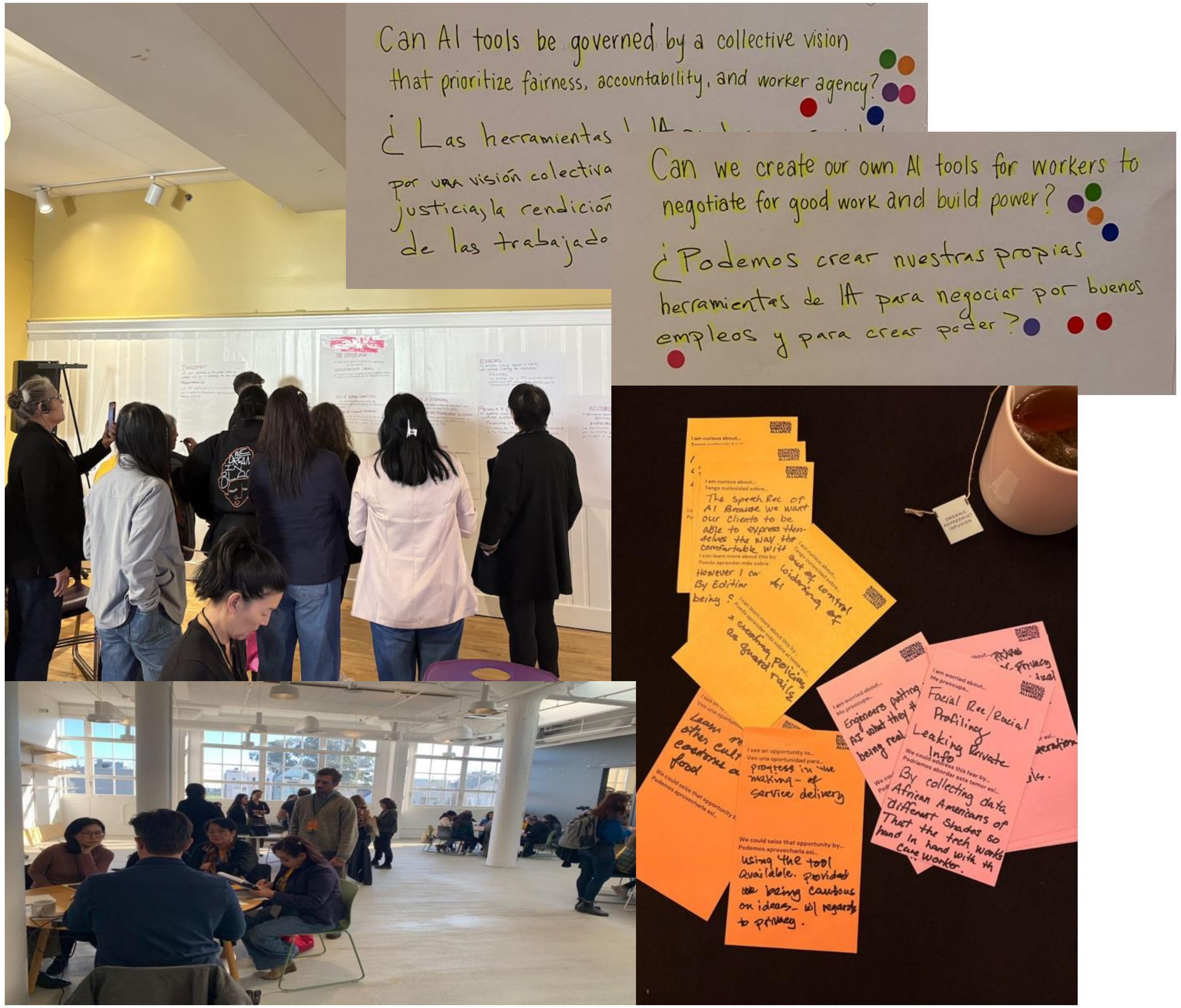
- Domestic care workers go into their clients' homes to take care of them.
    - Clients are often children, older adults, or people with disabilities
- This workforce is distributed
    - Enabled by gig work platforms
- Workers are often required to use AI tools
    - Tools can break in practice causing severe consequences
    - Tools can violate workers' and clients' privacy

There is a gap between the people building the technology and the people using it

Community partnerships

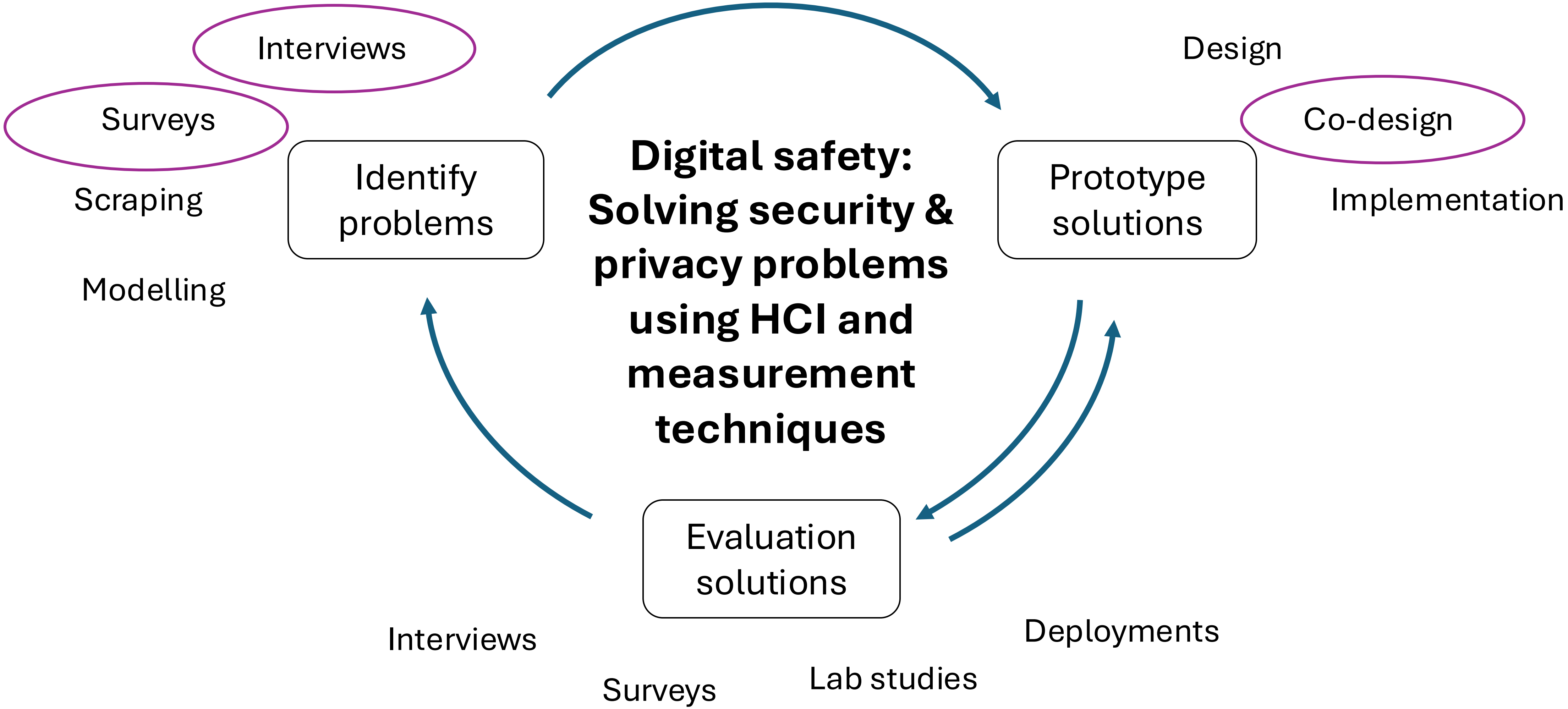
- Lay users are impacted by the technology we build, but have little power over it.
- Community partnerships give groups of users a greater voice

- **Goal:** close the gap between domestic workers and AI developers in care work
- **Example:**
  - 2 day workshop in SF
  - Bilingual curriculum pre-event


Can AI tools be governed by a collective vision that prioritize fairness, accountability, and worker agency?
¿Las herramientas
por una visión colectiva
justicia, la condición
de las trabajado
Can we create our own AI tools for workers to negotiate for good work and build power?
¿Podemos crear nuestras propias herramientas de IA para negociar por buenos empleos y para crear poder?
I am curious about...
Tengo curiosidad sobre...
The speech Rec of AI Because we want our Clients to be able to express them-selves the way the comfortable with
I can learn more about this by
Puedo aprender más sobre
However I ca
By Editin
being
out of control widening of AI
creating Policies as guard rails
I see an opportunity to...
Veo una oportunidad para...
progress in the making - of service delivery
We could seize that opportunity by...
Podemos aprovecharla así...
using the tool available. provided being cautious on ideas - w/ regards to privacy.
Learn other cult costome food
I am worried about...
Me preocupa...
Engineers putting AI what they being real
Facial Rec/Racial Profiling
Leaking Private Info
We could address this fear by...
Podríamos abordar este temor así...
By collecting data African Americans of different Shades so That the tech works hand In hand with the care worker.
ORGANIC PEPPERMINT INFUSION

Interested in chatting? Email me! varivera@cs.stanford.edu

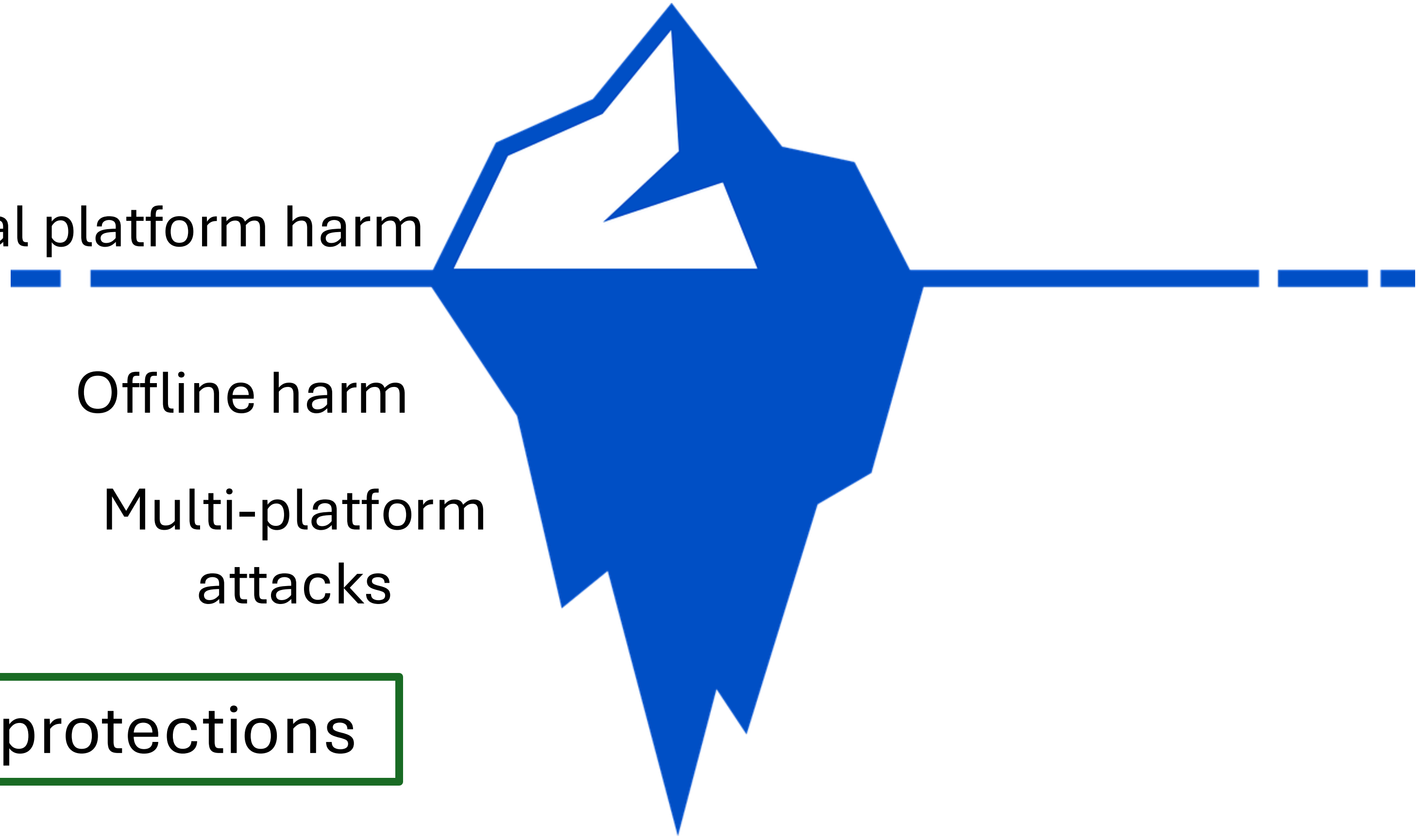
Want to do more research in security? Consider working with me and/or Alex at Georgia Tech's School of Cybersecurity and Privacy!



Digital safety: Solving security & privacy problems using HCI and measurement techniques

My vision: digital safety across an ecosystem

Current: Single platform protections



Individual platform harm

Offline harm

Multi-platform attacks

Future: Ecosystem-level protections

My vision: digital safety across an ecosystem

My approach:

- Problems from **security & privacy**
- Tools from **human-computer interaction**

Today's talk:

1. Algorithmically-mediated offline introductions
2. Bias & harassment in gig-work
3. Goals & threats in abuse reporting

Veronica Rivera Stanford University varivera@stanford.edu

Digital safety across an entire ecosystem

Solving digital safety problems with security techniques

Giving users greater agency over black box systems

Developing tools and techniques to measure harm at scale

Designing safer systems for abuse reporting

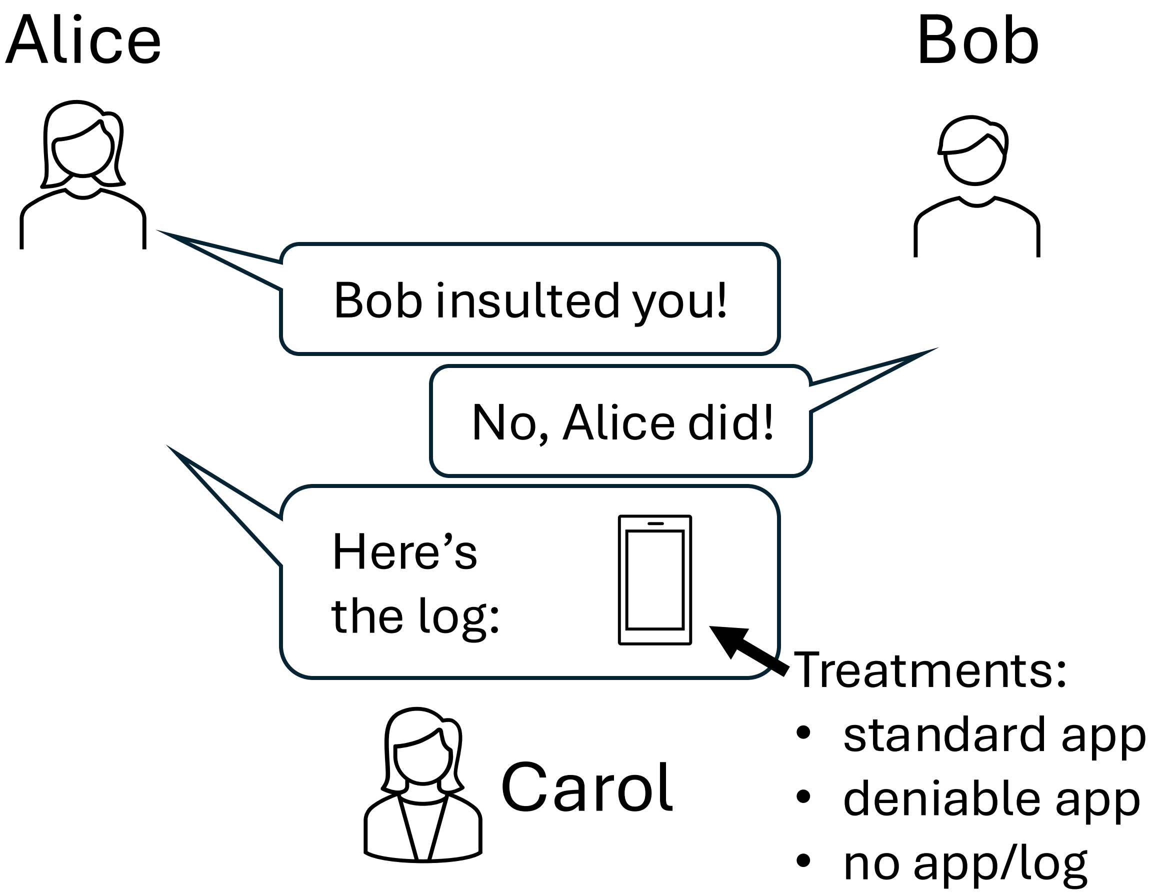
- How can we make peer-to-peer reporting safer?

- Idea: Use cryptographic tools such as **deniable encryption** and **secure reputation systems**

- Research question: Are these tools effective in practice?

Goal: **practical deniability**

- How to discuss sensitive topics **without retaliation**?
  - Chat logs ⇒ evidence?
- Connection: deniable encryption.
  - Theoretical cryptography
- Idea: **behavioral experiments**
  - ***human practicality*** of deniability



Alice

Bob

Bob insulted you!

No, Alice did!

Here's the log:

Carol

Treatments:
- standard app
- deniable app
- no app/log

Q: Who does Carol believe?

[Canetti, Dwork, Naor, Ostrovsky '97]

User-value alignment in LLM applications

We can't give users strong security guarantees over many AI systems. Can we give them greater agency?

Research questions:

- How can users shape the outputs of LLMs to better align with their definitions of safety?
- How can training and building AI be more participatory in high-stakes deployments?

Measuring harm: a foundation for empiricism

To know whether we’ve reduced harm overall, we must be able to measure it at scale

Research questions:

- How do we collect statistics while preserving individual privacy?
- Who do users trust to collect such statistics?
- How do we measure harm at scale: across platforms and user groups?

Impact in research, tool development, and policy

Charting new research directions

Building digital safety tools

NIST
National Institute of Standards and Technology

Guiding policy for AI safety*

* Response to NIST's AI Executive Order Request for Information published via Georgetown University's Massive Data Institute. By Elissa M. Redmiles, Sarah Adel Bargal, Grace (Natalie) Brigham, Nina Grgic-Hlaca, Tadayoshi Kohno, Jaron Mink, **Veronica A. Rivera**, Carmela Troncoso, Lucy Qin, Miranda Wei

Focus: Digital whisper networks for **labor abuse**

Labor abuse includes:

- Physical violence
- Harassment
- Scams
- Wage theft
- Plagiarism of work by colleagues