Safety in Algorithmically-Mediated Offline Introductions: Lessons for Research, Design, and Policy

Veronica A. Rivera, Ph.D

Postdoctoral Researcher, Stanford University

varivera@stanford.edu

Work with Daricia Wilkinson (Microsoft Research), Aurelia Augusta (Max Planck Institute), Sophie Li (Max Planck Institute), Elissa M. Redmiles (Georgetown), and Angelika Strohmayer (Northumbria)

Stanford University
Human-Centered
Artificial Intelligence

Stanford | McCoy Family Center for Ethics in Society

About me:

Stanford Empirical Security Research Group

Research on digital safety and security of marginalized and vulnerable populations

HARVEY MUDD COLLEGE

BS in Computer Science and Math

UC SANTA CRUZ

PhD in Computational Media (HCI)

MAX PLANCK INSTITUTE FOR SOFTWARE SYSTEMS

PRISM
Center for Privacy and Security of Marginalized and Vulnerable Populations

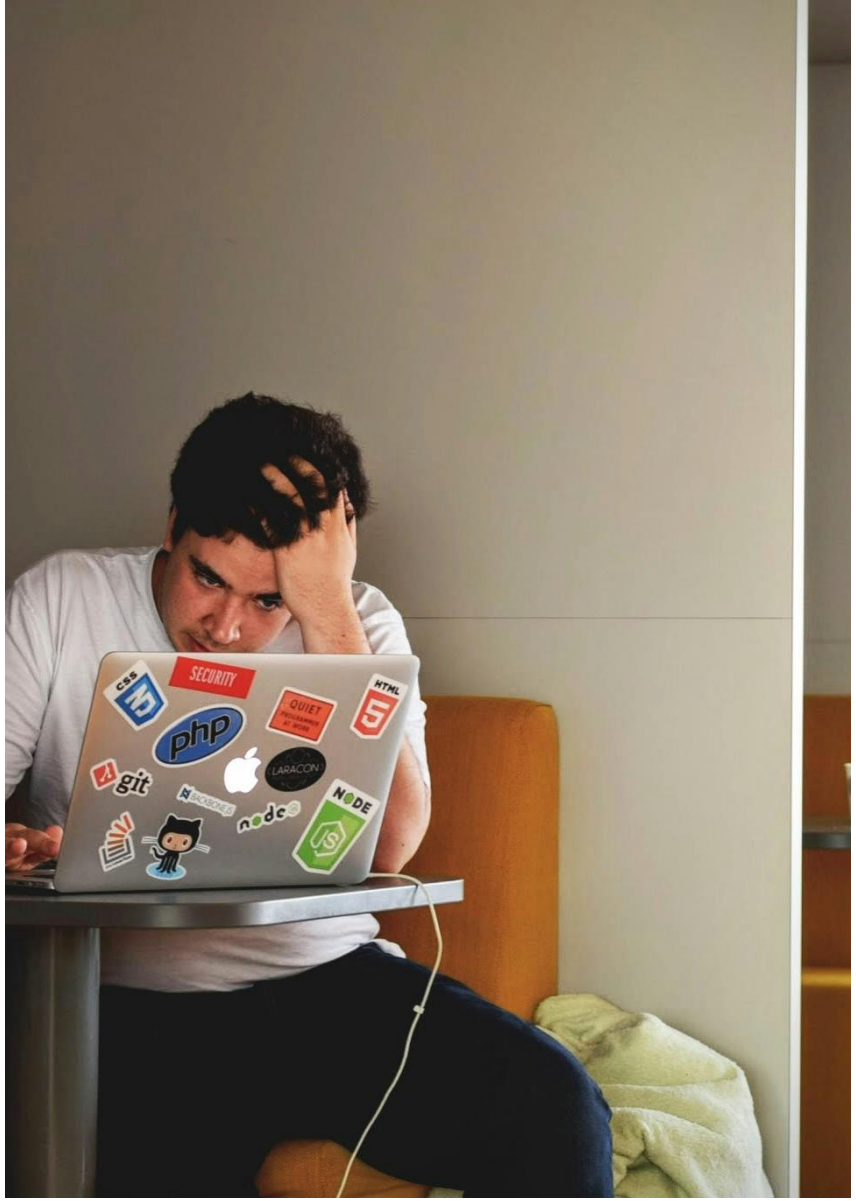
Visiting PhD student at the Max Planck Institute for Software Systems (Germany) & at Center for Privacy and Security of Marginalized and Vulnerable Populations (UF)

HAI
Stanford University
Human-Centered Artificial Intelligence

Stanford | McCoy Family Center for Ethics in Society
SCHOOL OF HUMANITIES & SCIENCES

Working with philosophers to create ethics curriculum for undergraduate CS courses

Experiencing online hate and harassment **impacts an individual's mental health**



Tech-facilitated violence **impacts an individual's offline relationships**

Online platforms can **facilitate physical abuse**

Digital harm can cross the online-offline divide

Online stalking

Doxing

Non-consensual sharing

Benign conversation online

Mental health

Offline relationships

Physical abuse

Digital harm is impacted by many components

Goals

Learning a skill

Finding work

Meeting others

Sharing major life events

Talking with friends

Gaming

Actions

Posting a photo or video

"Friending" someone

Viewing a photo

Joining a group

Clicking a link

Uploading information

Harms

Scams

Doxing

Emotional abuse

Harassment

Physical violence

Financial abuse

Misinformation

Non-consensual sharing

Risk Factors

Gender

Socio-economic

Power

Internet skill

Social expectations

Disability

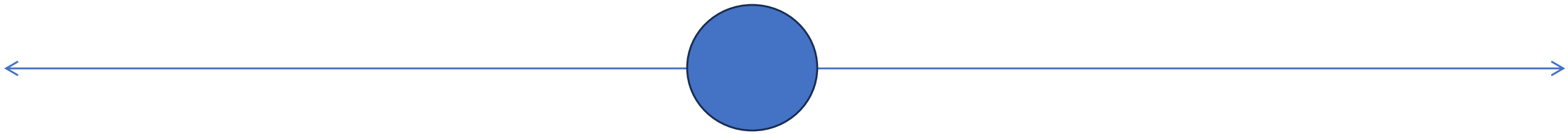
Race

Education

The complexity of digital harm makes protection challenging

Overly generalized approaches

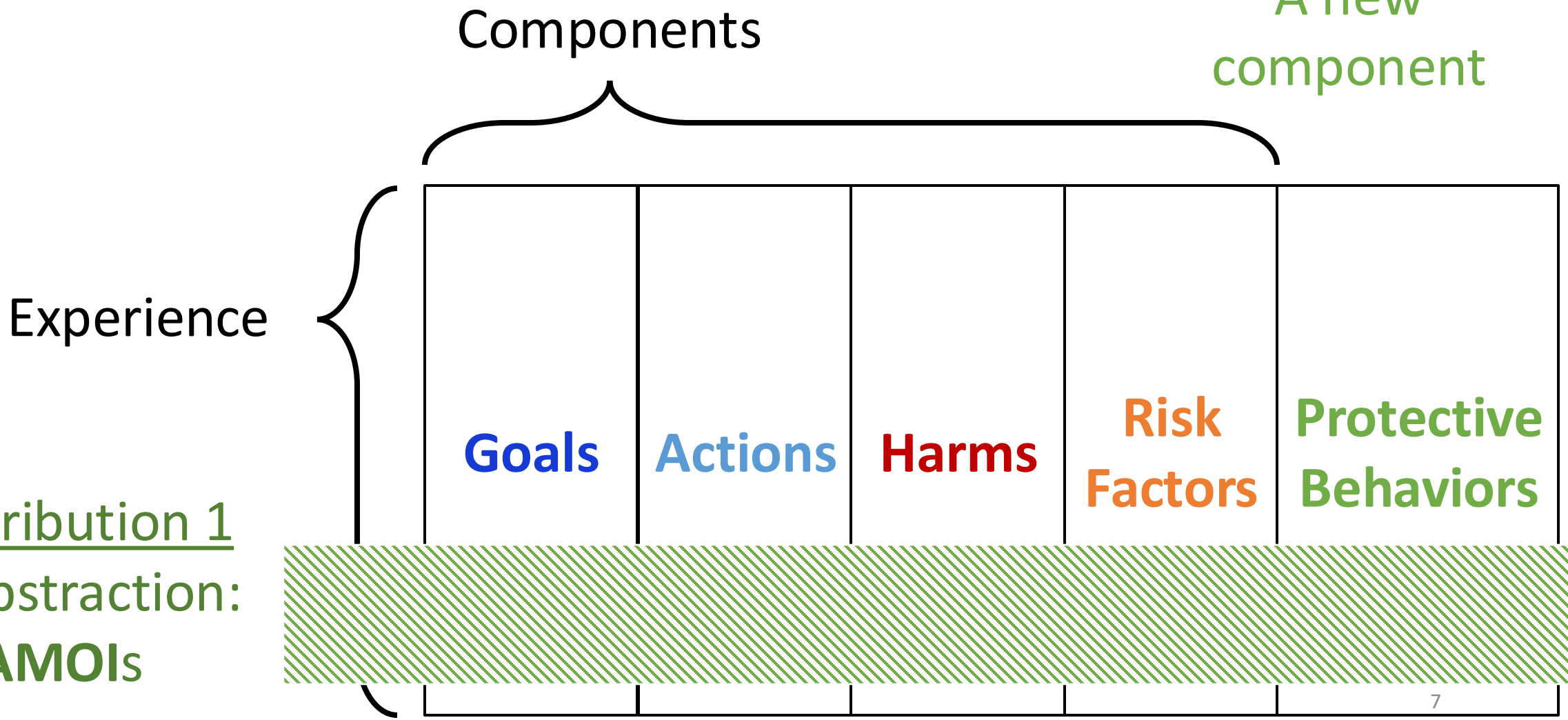
Mitigation for specific populations



Safety: experiences and components

Contribution 2

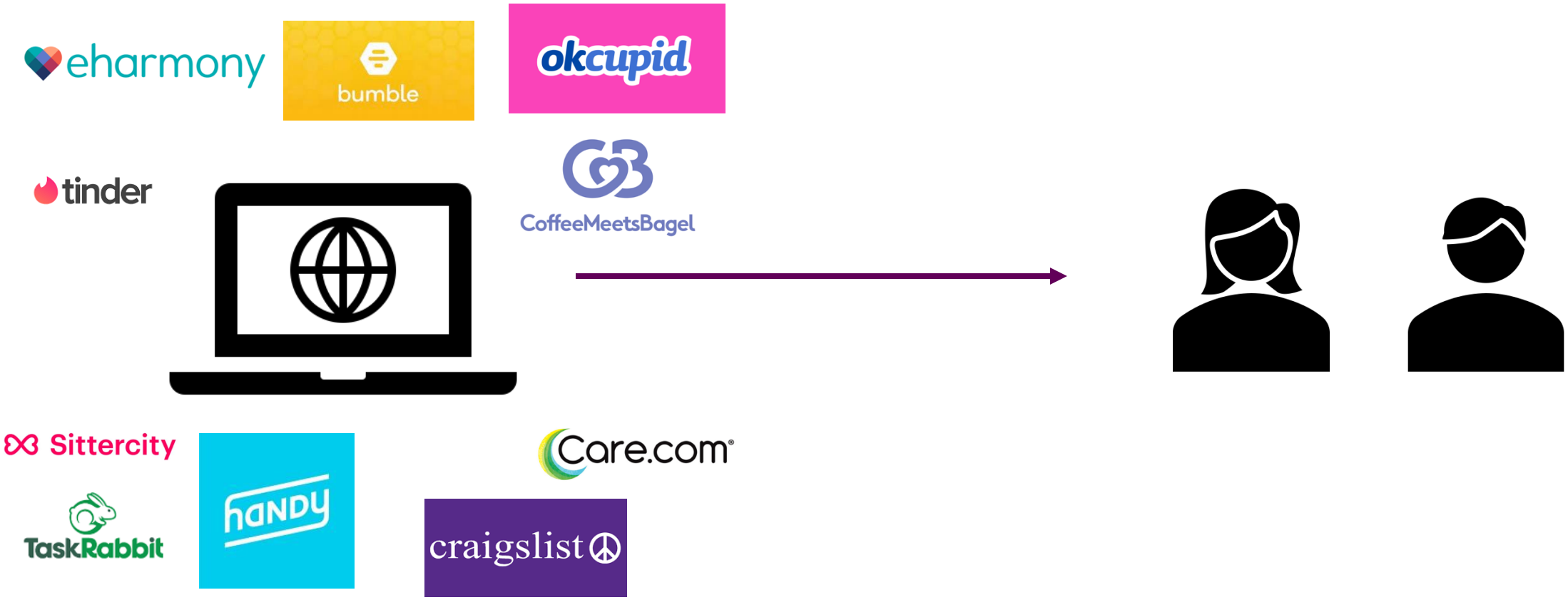
A new component

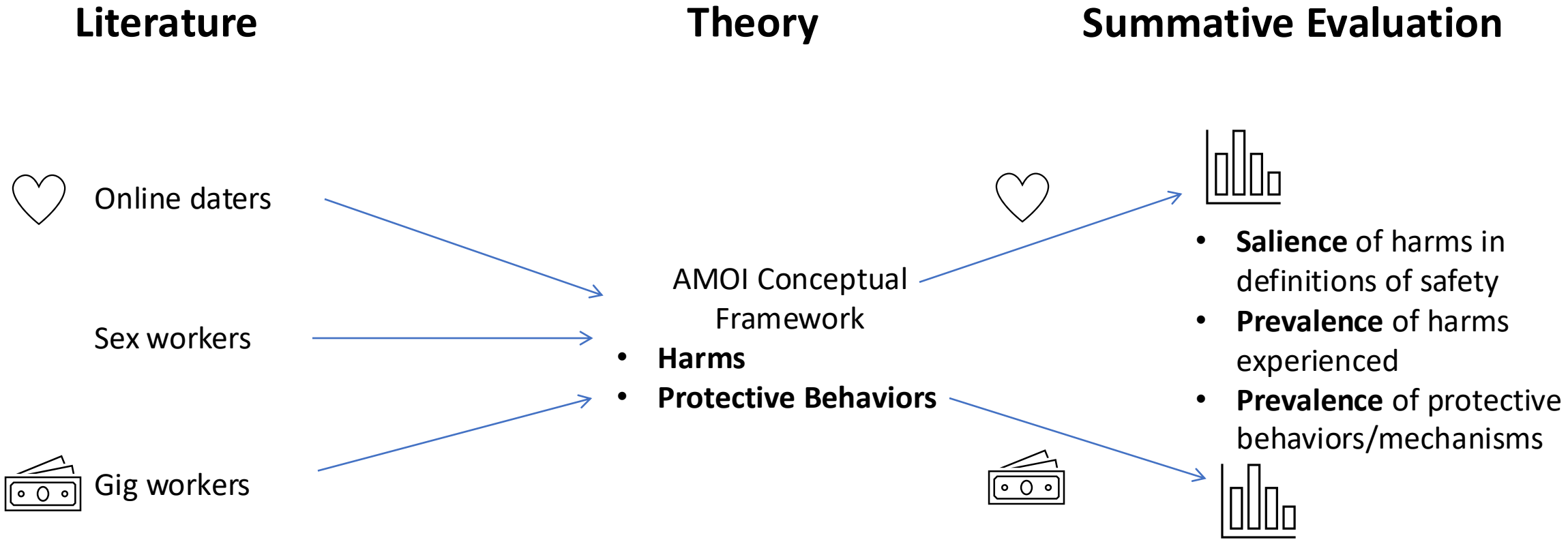


Contribution 1

An abstraction: **AMOI**s

Algorithmically-mediated offline introduction (AMOI): an offline introduction between strangers that is mediated by a matching algorithm on an app or website





Veronica A. Rivera, Daricia Wilkinson, Aurelia Augusta, Sophie Li, Elissa M. Redmiles, Angelika Strohmayer. Safer Algorithmically-Mediated Offline Introductions: Harms and Protective Behaviors. *To appear at the Conference on Computer-Supported Cooperative Work and Social Computing (CSCW) 2024*

Roadmap

- Study objectives & methods
- Part 2: Harms
  - Conceptual framework
  - Insights from survey data
- Part 3: Protective behaviors
  - Conceptual framework
  - Insights from survey data
- Part 4: Takeaways for future work
  - Tech design
  - Policy

Part 1: Study objectives and methods

Mixed methods: systematic literature review & survey

Systematic lit review: Mapping an area

A systematic literature review is a research method used to obtain and evaluate a corpus of research articles to answer a research question.

Keywords:

- Example: “gig work” and “online dating” and “sex work” + “safety”, “harm”, “scam”, “security”

Databases queried:

- Google Scholar
- ACM Digital Library
- ScienceDirect
- Springer Link
- IEEE Xplore Digital Library

Process:

- Reviewed titles, abstract, and conclusion for relevance, focusing on whether they discuss
  - Harms
  - Protective behaviors
  - Resources/mechanisms by which people carry out those behaviors

Summative evaluation: Measuring harms & behaviors

Participants

	Online Daters	Gig Workers
Lucid	104	
Prolific	372	451

Process

- Duration: Aug’21 to Dec’21
- Stages:
  - Screener Survey; criteria:
    - US-based
    - Used an app within 2 years
  - Main Survey
- Gender & race matched US census

Survey contents

1. How they **define safety**
   (single open-response question)

2. What **harm** they have experienced
   (single multiple response questions)

3. What **protective behaviors** they engage in and
   what **resources/mechanisms** are used to carry them out
   (several multiple response questions)

Part 2: Harms

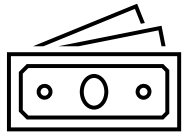
Goal

Systematize the **harms** prior work has identified and **measure the salience of those harms** in people's definitions of safety

Harms conceptual framework (literature)

Physical

- Bodily harm from assault, abuse, and/or disease



Financial

- Fraud, internet scams
- Financial instability arising from these

Privacy

- Misuse & abuse of personal information provided to platforms (e.g. for surveillance; by a malicious actor)

Autonomy

- Controlling what someone can and cannot do (e.g., by platforms or people)
- Lack of transparency over platforms' use of personal data

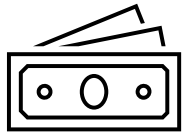
Emotional

- Consequence of other harms
- Fear over future harm

Harms in AMOI: adding contextual nuance

Physical

- Bodily harm from assault, abuse, and/or disease



Financial

”[Safety means] that I do not get harmed or robbed while being out…”

Privacy

- Misuse & abuse of personal information provided to platforms (e.g. for surveillance; by a malicious actor)

Autonomy

“It is important to always have an escape plan and **ensure you don’t get stuck**”

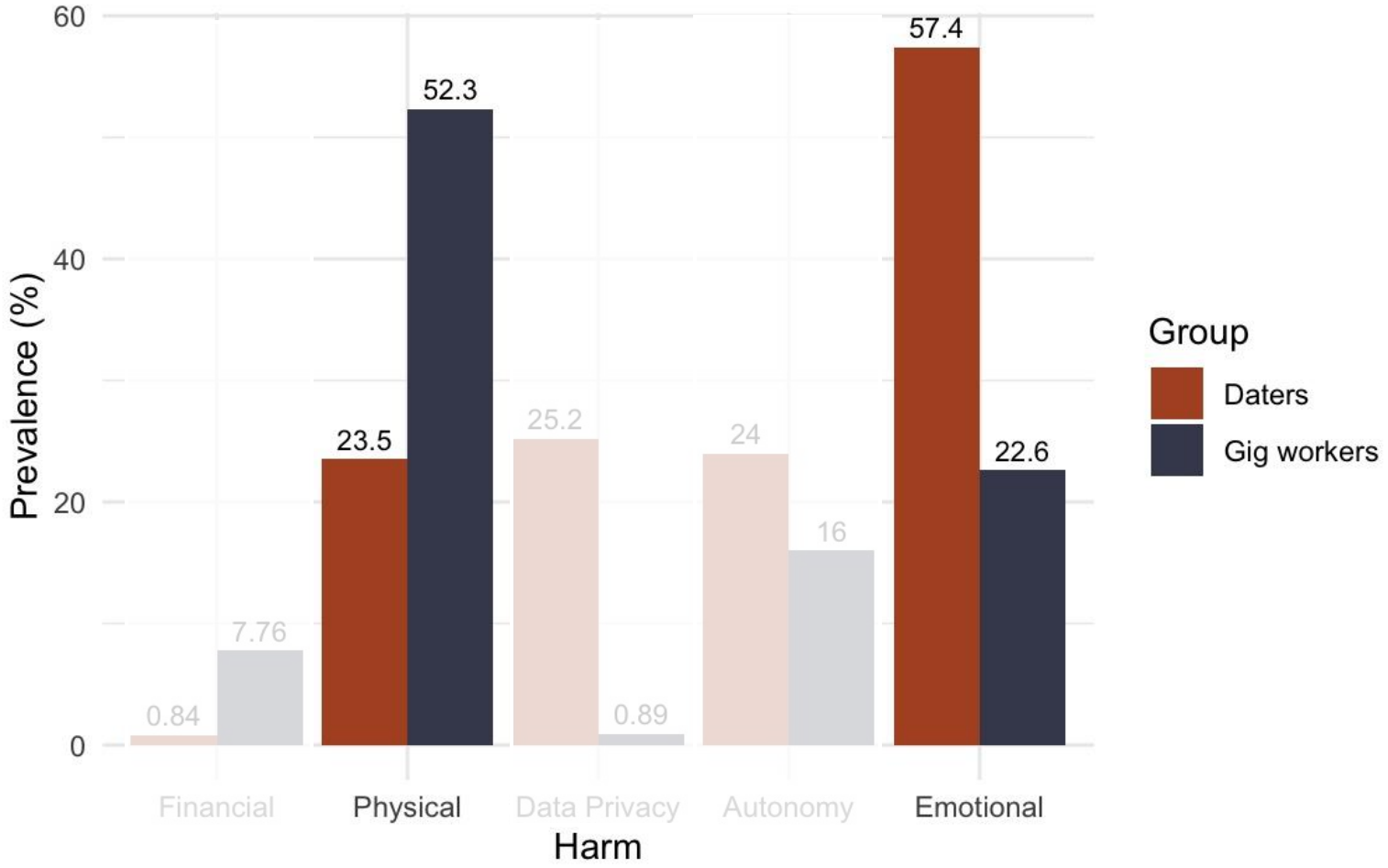
Emotional

- Consequence of other harms
- Fear over future harm

Physical & emotional harm are most salient in definitions of safety

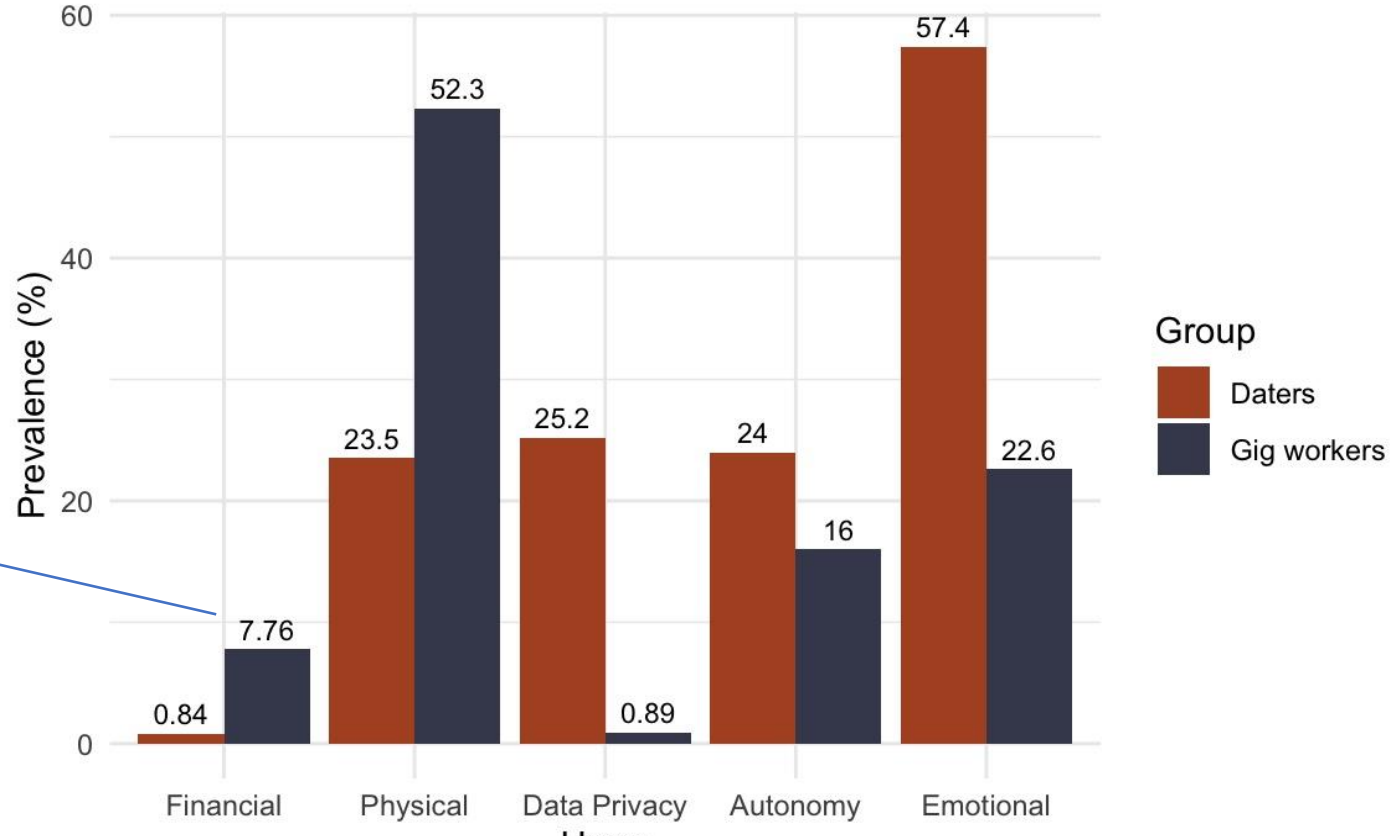
Emotional harm is the most salient concern in online daters' definitions of safety (57.4%)

Physical harm is the most salient concern in gig workers' definitions of safety (52.3 %)



Harm: definitional salience ≠ experienced frequency

Experienced by 41%

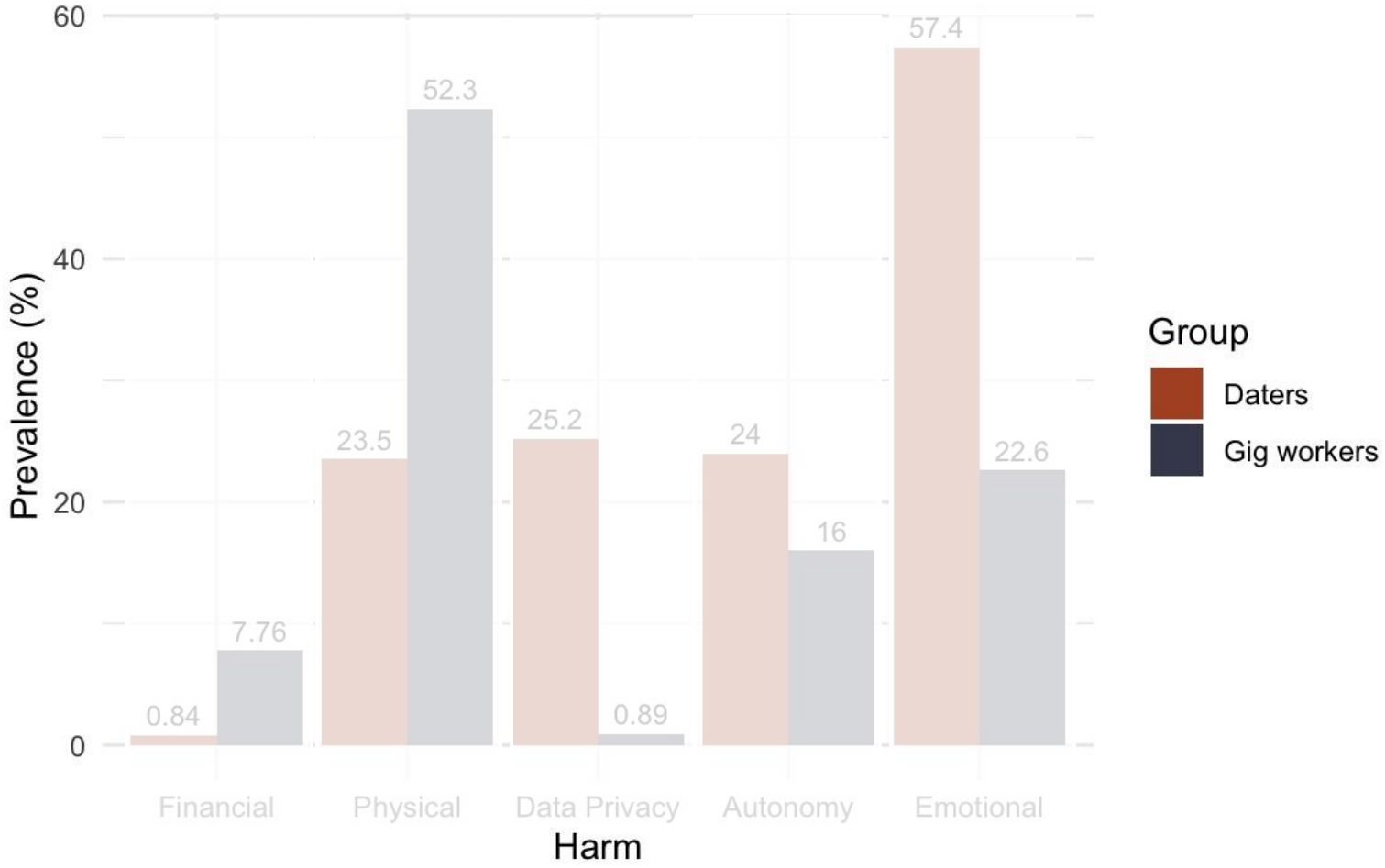


Salience of harms in participant **definitions**

Misalignment in what harms are prioritized in research

Financial harm overly focused on in both online dating and gig work literature relative to more salient harms

Autonomy harm under focused on in online dating literature

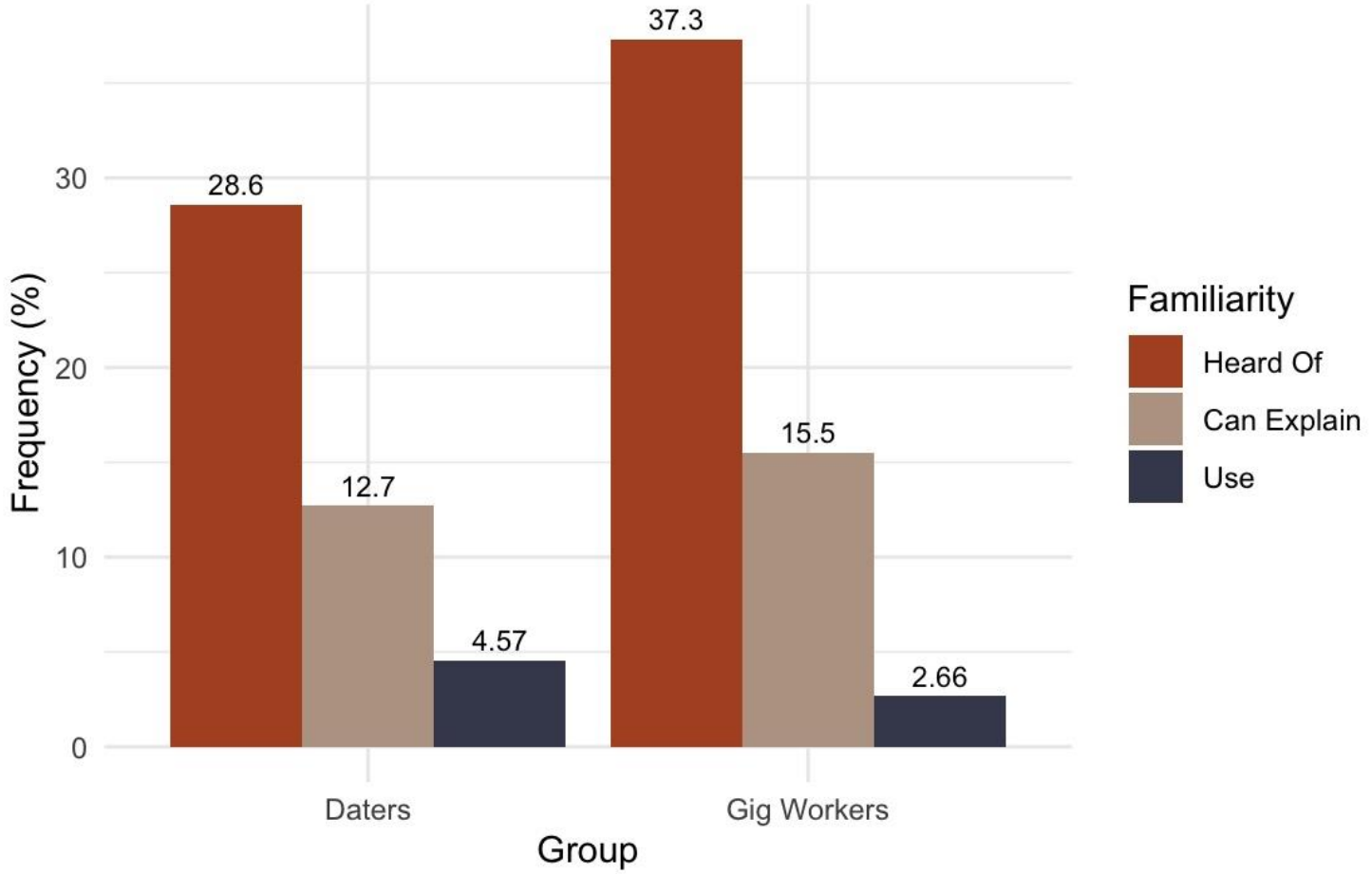


Part 3: Protective behaviors

Goal

Systematize the **protective behaviors** prior work has identified and measure the **prevalence of adoption of those behaviors** and the **resources/mechanisms** used to carry them out

People rarely use dedicated safety tools



Behavior mechanisms (literature)

Platform

Individual

Social

Screening

Self-disclosure

Obfuscation

Reporting

Blocking

Environmental precautions

Emergency alerts

Surveillance & documentation

Vetting

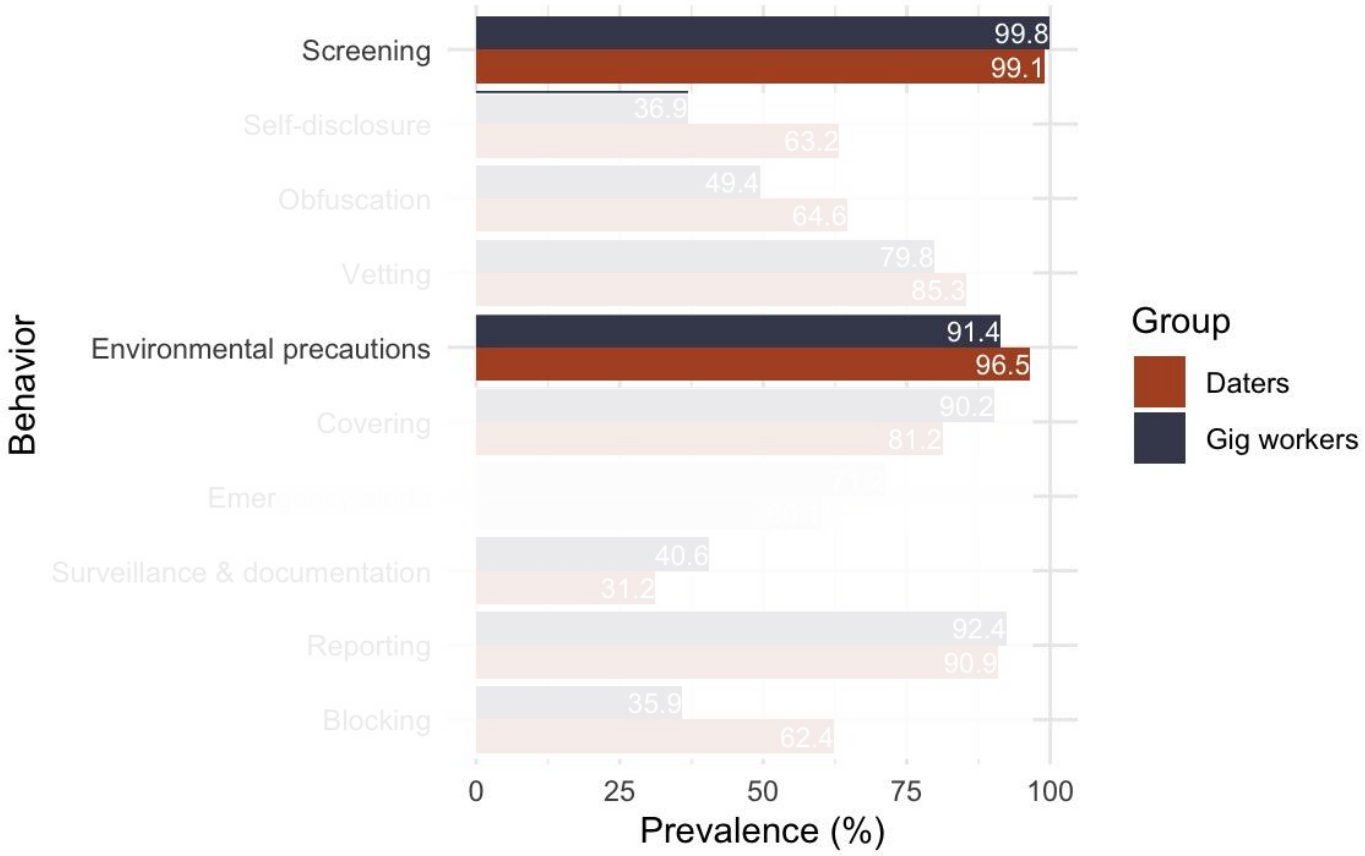
Covering

Reporting

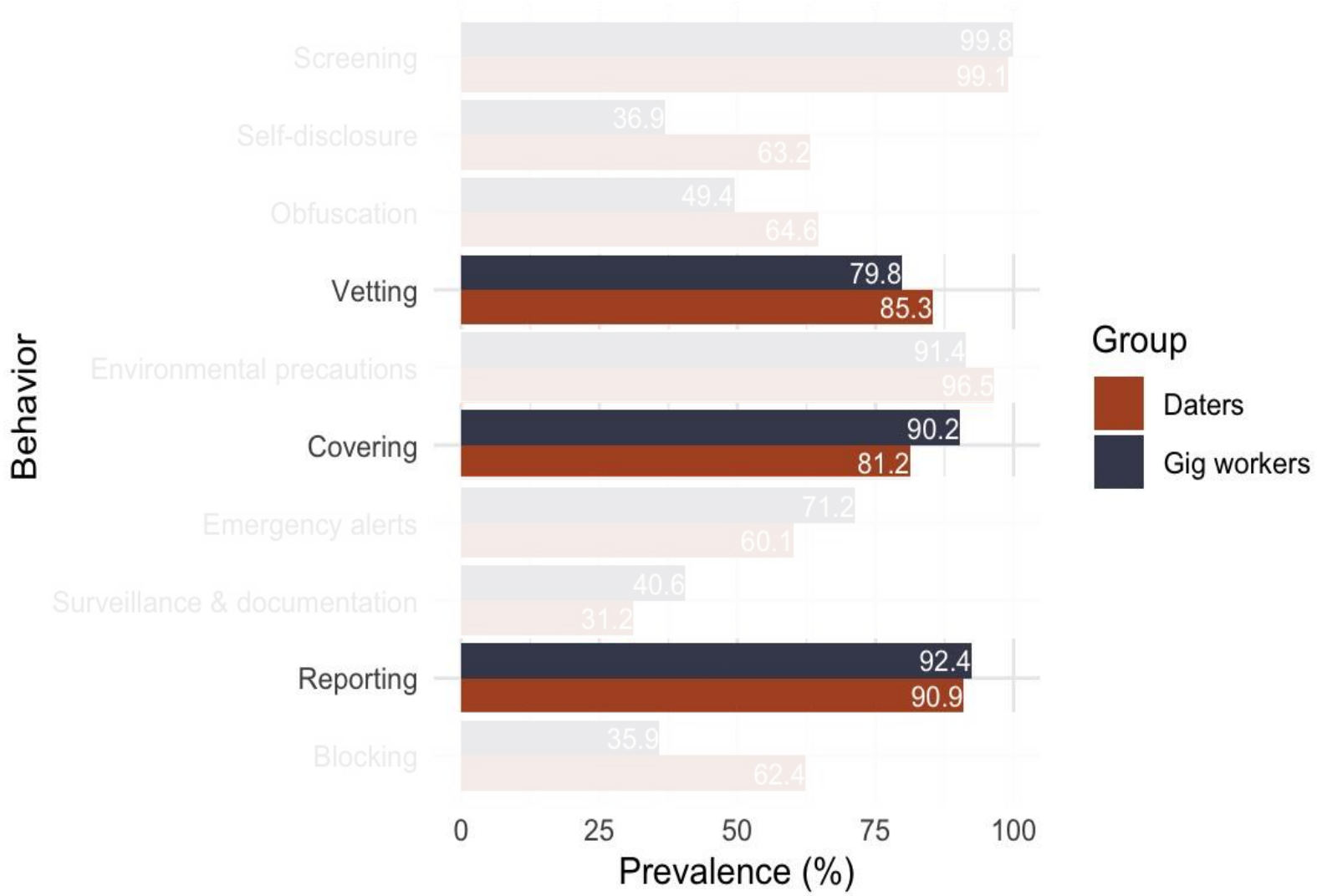
Individual behaviors that are easy to carry out are almost always used

~ 99% of respondents in both groups **screen** the people they will meet offline

> 95% of online daters & > 90% of gig workers engage in **environmental precautions**



Some of the most used behaviors are social

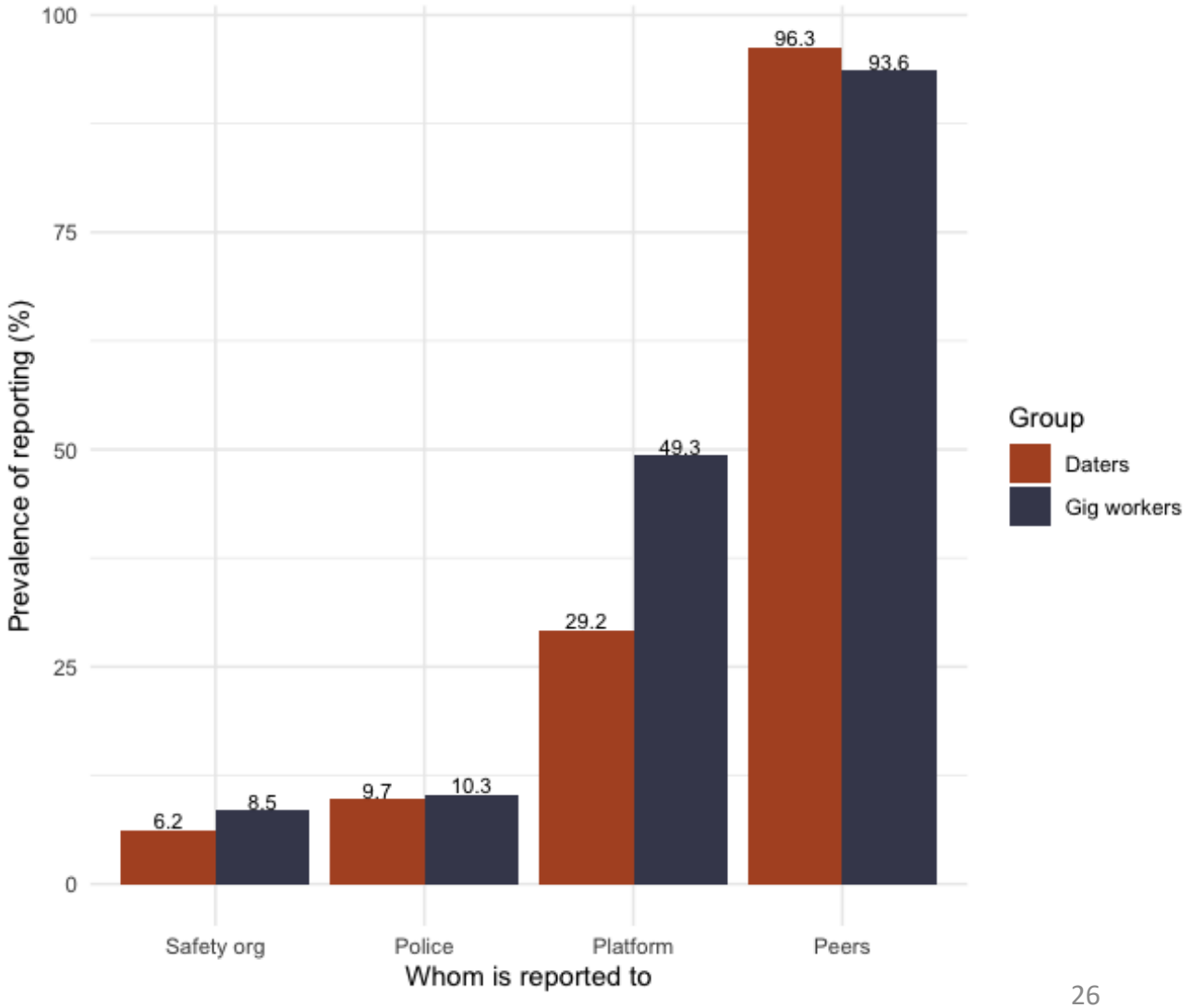


Reporting harm is most commonly a social behavior

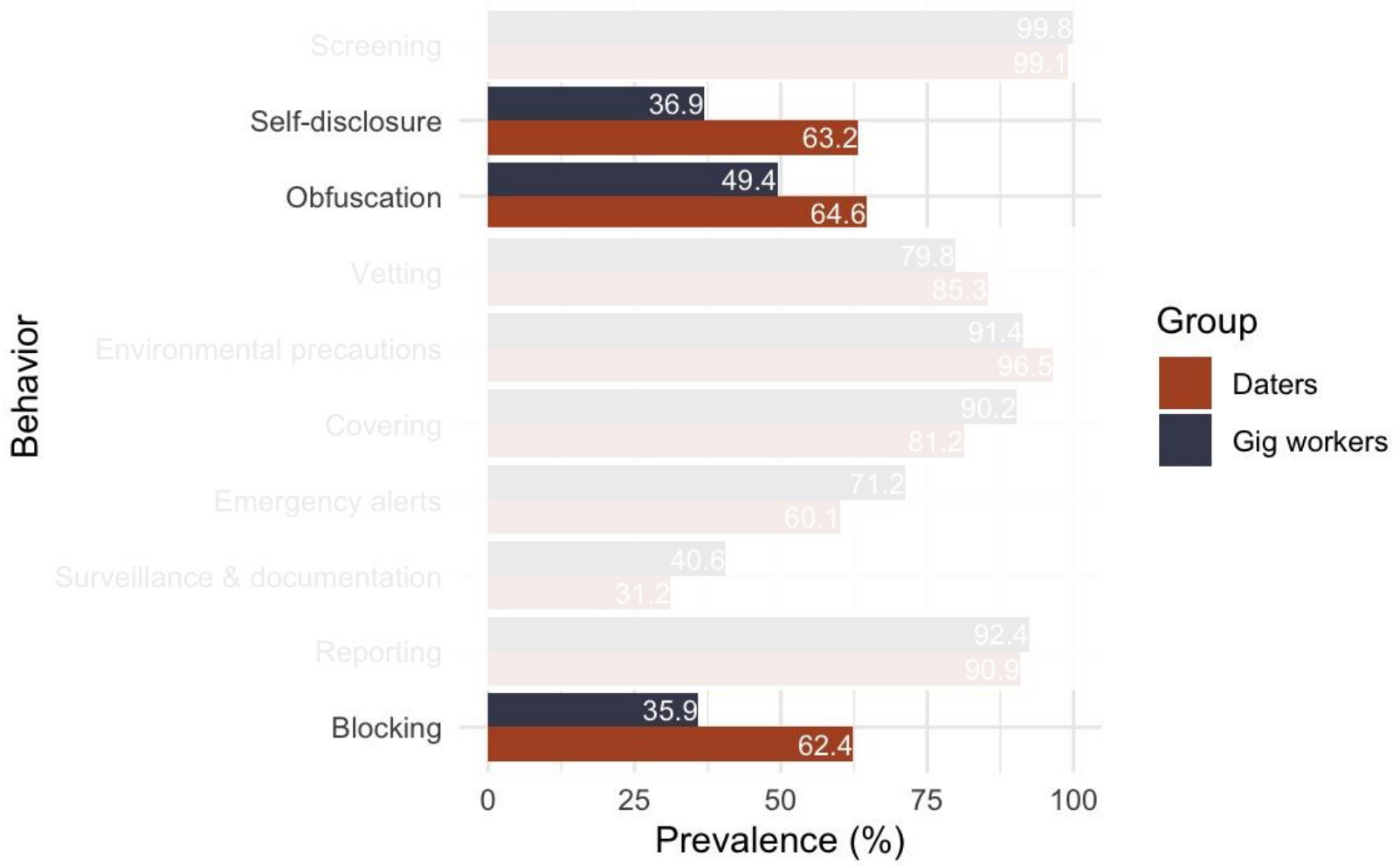
< 30% of online daters and < 50% of gig workers have reported harm to **platforms**

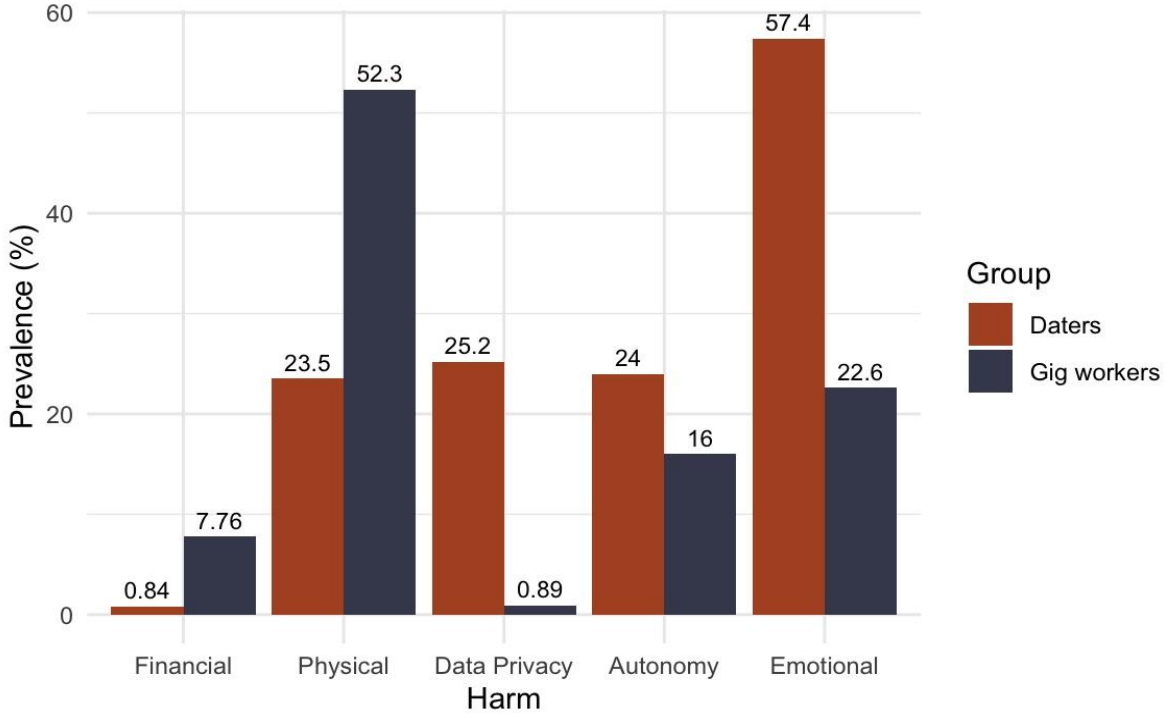
~ 10% of online daters and gig workers have reported harm to **law enforcement**

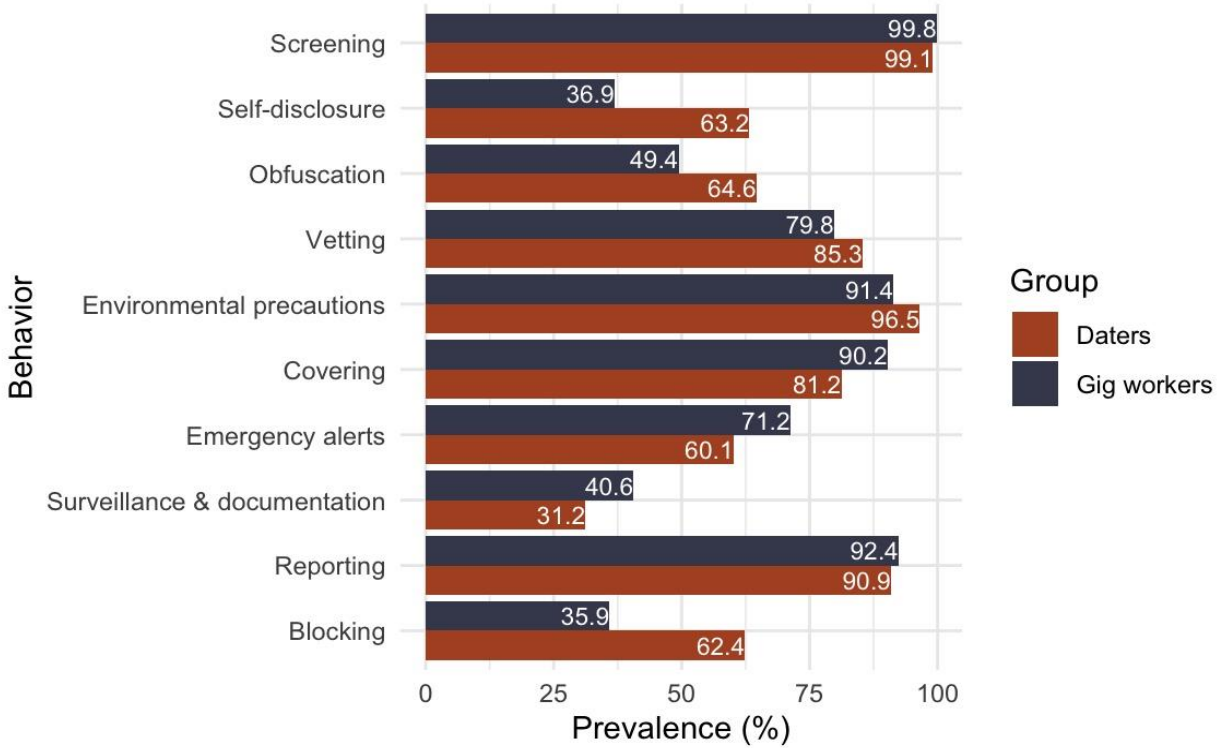
< 10% of online daters and gig workers have reported harm to **safety NGOs**



Some behaviors rely on platform design







In AMOIs users are vulnerable but not helpless.
Perhaps we can learn from their protective behaviors to inform the design of future safety mitigations

Policy translation

White House Gender Policy Council

NIST
National Institute of Standards and Technology

Best practices for AI safety*

Formal 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

Part 4: AMOI applications to Tech-Facilitated Gender-Based Violence

With Hanna Barakat and Elissa M. Redmiles

What is Tech-Facilitated Gender-Based Violence (TFGBV)

The UN defines TFGBV as "any act that is committed or amplified using digital tools or technologies causing physical, sexual, psychological, social, political, or economic harm to women and girls because of their gender."

- Online harassment
- Revealing someone's personal info without their consent
- Intimate image abuse
- Sharing deepfake images that contain someone's face and/or body
- Stalking
- Physical violence

We approach TFGBV through a broader lens

- Violence against marginalized and/or vulnerable groups
- Harm that amplifies existing inequalities
- Is *enabled by* or *mediated through* digital technologies.

Existing TFGBV policy landscape

- **US:** In May 2023 the White House released a National Plan to End Gender-Based Violence across urban, suburban, rural, and Tribal communities in the US.
- **EU:** The Digital Services Act and the Online Safety Act (UK) both aim to increase accountability for harm caused on digital platforms.
- **Australia:** The eSafety Commissioner, the world's first government online safety regulator, is working to make digital spaces safer for women and promote greater gender equity.

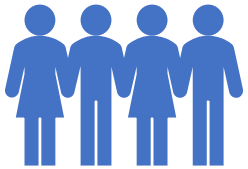
Challenges in addressing TFGBV

TFGBV & AMOI parallels

- Harm crosses the digital-physical divide
- Harm is affected by individual risk factors and tech use
- Lack of clarity around who's responsible for harms that impact life beyond the platform

Our recommendations for addressing TFGBV

Harm reduction

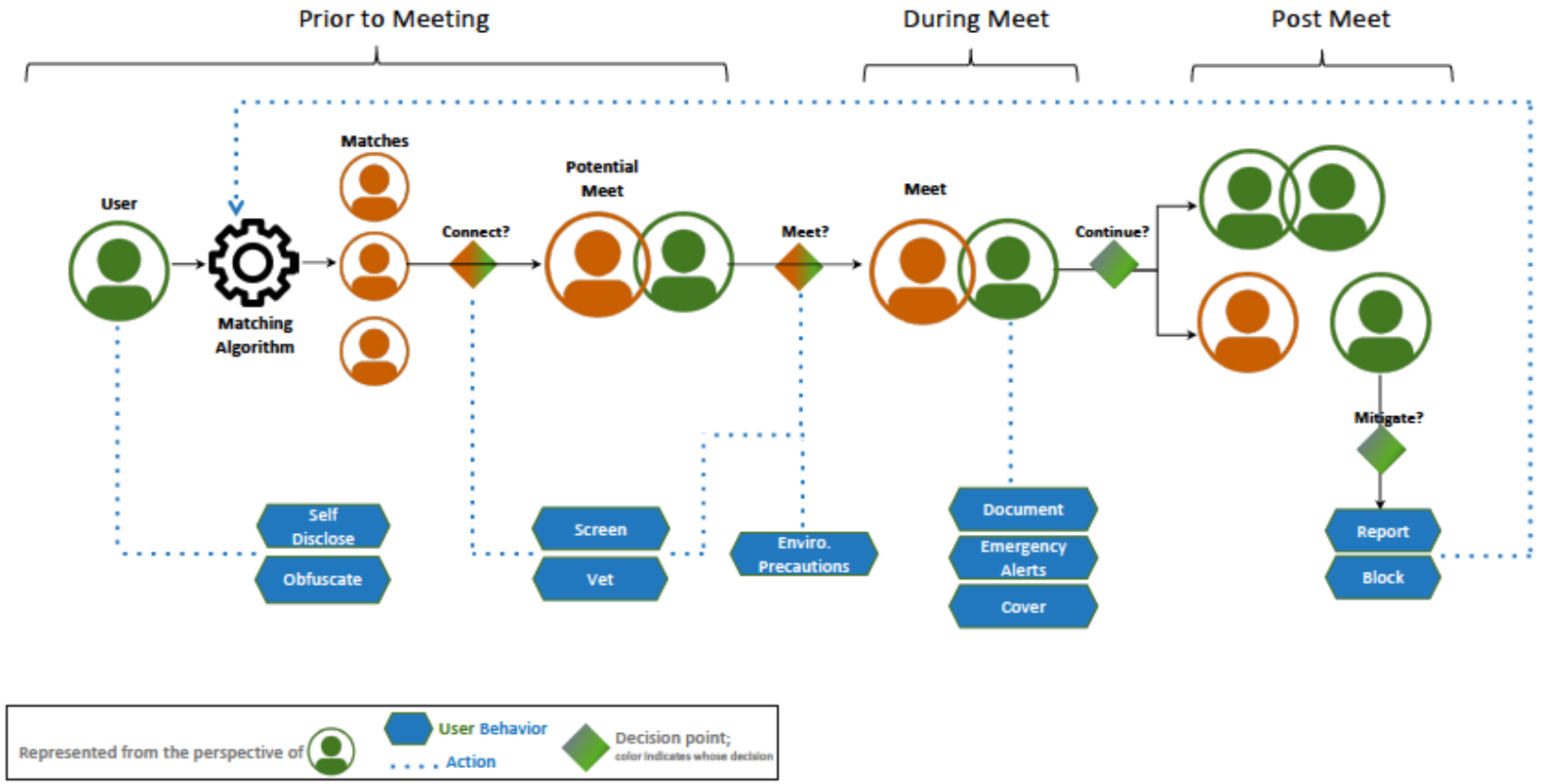


Survivor support

Platform accountability

Harm reduction

Recommendation 1: Leverage threat models to proactively identify and address where harm will occur [tech]



Harm reduction

Recommendation 2: Establish protections that address ongoing harm in addition to isolated instances [policy]

- Establish policies that require platforms to implement survivor-led reporting mechanisms to indicate recurring abuse
- Monitor patterns in chronic harm

Survivor support

Recommendation 3: Increase transparency and usability of reporting systems [policymakers, platforms, civil society]

- Platforms should implement clearly identified reporting mechanisms which are overseen by regulators
- NGOs could independently implement cross-platform reporting mechanisms to allow people to experience harm across multiple platforms
  - Australia's eSafety commissioner does something like this
- Privacy-preserving reports could be publicly made available to increase user transparency into harms

Accountability

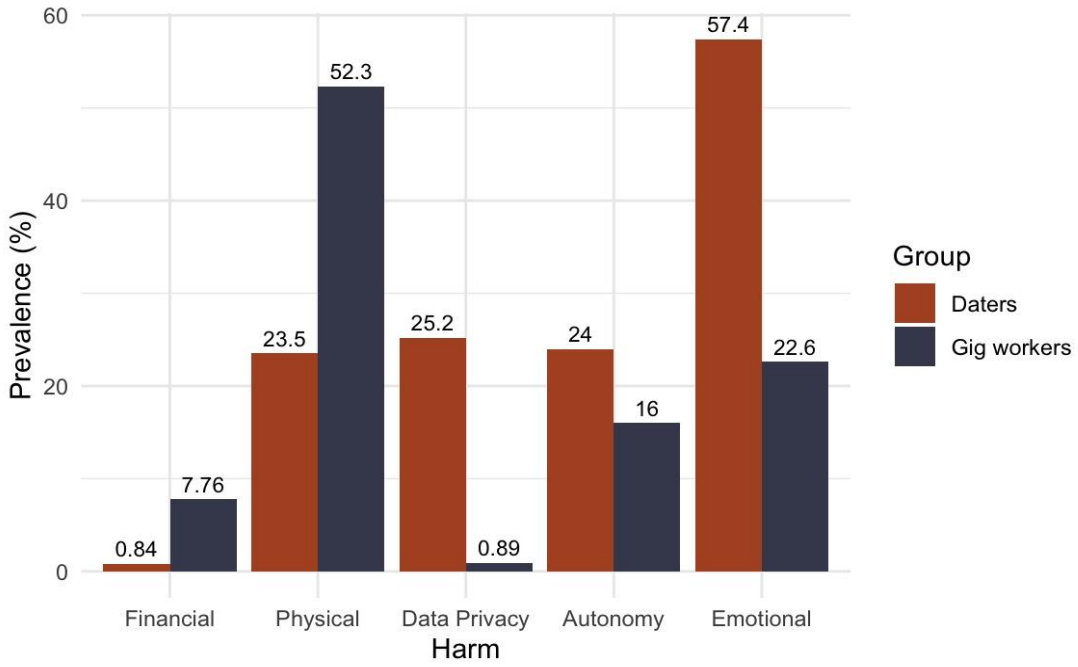
Recommendation 4: Survey survivors and at-risk individuals to understand their trust in different entities to guide safety strategies

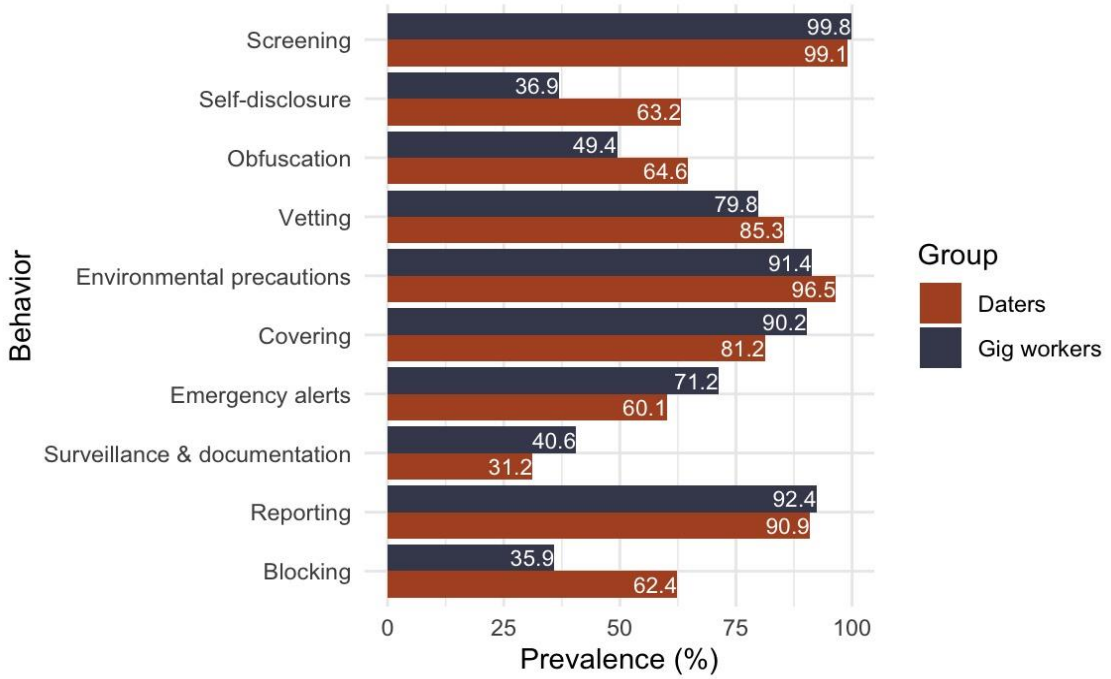
[academia, policymakers, civil society]

- Conduct large-scale surveys to evaluate the perceived trustworthiness of various institutions to address harm at different points in time
- Use the results of this work to inform what kind of support different stakeholders are best positioned to offer

Safety in Algorithmically-Mediated Offline Introductions

Veronica A. Rivera
varivera@stanford.edu
https://vrivera2017.github.io





The right abstractions across experiences can provide insight for addressing digital harm from a tech and policy lens