

“Just Like, Risking Your Life Here”: Participatory Design of User Interactions with Risk Detection AI to Prevent Online-to-Offline Harm Through Dating Apps

ISHA DATEY, Oakland University, USA

DOUGLAS ZYTKO, University of Michigan-Flint, USA

Social computing platforms facilitate interpersonal harms that manifest across online and physical realms such as sexual violence between online daters and sexual grooming through social media. Risk detection AI has emerged as an approach to preventing such harms, however a myopic focus on computational performance has been criticized in HCI literature for failing to consider how users should interact with risk detection AI to stay safe. In this paper we report an interview study with woman-identifying online daters (n=20) about how they envision interacting with risk detection AI and how risk detection models can be designed pursuant to such interactions. In accordance with this goal, we engaged women in risk detection model building exercises to build their own risk detection models. Findings show that women anticipate interacting with risk detection AI to augment - not replace - their personal risk assessment strategies. They likewise designed risk detection models to amplify their subjective and admittedly biased indicators of risk. Design implications involve the notion of personalizable risk detection models, but also ethical concerns around perpetuating problematic stereotypes associated with risk.

CCS Concepts: • **Human-centered computing** → **Empirical studies in HCI; User models**; • **Social and professional topics** → **Women**; • **General and reference** → **Empirical studies**; • **Computing methodologies** → **Artificial intelligence**.

Additional Key Words and Phrases: Computer mediated Harm, Risk detection, Participatory design, Women, Harm, Risk, Artificial Intelligence, Dating Apps

ACM Reference Format:

Isha Datey and Douglas Zytke. 2024. “Just Like, Risking Your Life Here”: Participatory Design of User Interactions with Risk Detection AI to Prevent Online-to-Offline Harm Through Dating Apps. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 367 (November 2024), 41 pages. <https://doi.org/10.1145/3686906>

1 Introduction

More than 800 million women have been victims of sexual or physical violence in their lifetime [95], and research consistently finds women to be at disproportionate risk of sexual harm [2, 15, 39, 95]. Mounting evidence demonstrates that social computing platforms are amplifying the risk of harm against women, not only online but in the physical world [38, 46, 60, 84, 104]. Of these, we focus on online-to-offline harm which refers to instances where harm, such as sexual violence or bodily injury, occurs in the physical world as a result of prior online communication. This includes harm such as use of social media to lure victims into physical sexual abuse [116] or sex trafficking [100, 129], and non-consensual sexual activity committed by someone met on a dating app [102, 140].

Authors' Contact Information: Isha Datey, Oakland University, Rochester, Michigan, USA, ishadatey@oakland.edu; Douglas Zytke, University of Michigan-Flint, Flint, Michigan, USA, dzytko@umich.edu.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

© 2024 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM 2573-0142/2024/11-ART367

<https://doi.org/10.1145/3686906>

Risk detection AI has emerged as a promising approach for scalable mitigation of online-to-offline harm [37, 100, 101]. Examples include detection of sex traffickers online [129], cybergrooming of children [14], and safety threats against women in urban environments [94]. Risk detection AI has also been applied to traditionally online harms such as unsolicited nude images on social media [121]. Despite a growing number of application contexts, the HCI literature has identified several concerns with the current state of risk detection AI that necessitate involvement of anticipated users in design and development [59, 69, 71, 83, 100, 118, 120]. First are issues with the AI's technical capabilities to detect risk: there is a lack of ecologically valid datasets for model training that represent real experiences of computer-mediated harm [100], coupled with an over-reliance on external annotators who may never have personally experienced the harms they are labeling, thus contributing to inconsistencies in ground truth [101, 106, 139]. Stakeholders have been involved in addressing these issues through voluntary donation [16, 57, 99] and annotation [10, 41, 99] of their own data reflective of harm, such as private messaging interactions [10].

Second are gaps in knowledge about how to design human interaction with risk detection AI so that the detection of risk can have a preventative impact on user safety. Tariq and colleagues describe this most succinctly as the difference between "risk detection" and "risk mitigation" [120]. Prior research into computational performance of "after-the-fact" risk detection models [33, 40, 71] leave open questions of how risk detection AI should be embedded in the user experience to provide practical benefit. This goes beyond surface-level questions of interface design (e.g., how to explain a risk detection model's output) because anticipated interactions with risk detection AI can recursively inform the AI's development - what risk means to users, how they expect risk to be computed, when and how they are informed of risk, and so on.

We explore these gaps with the following research questions: **(RQ1)** *How do women anticipate interacting with or using risk detection AI to protect themselves from online-to-offline harm?* **(RQ2)** *How should online-to-offline risk detection AI be designed to fulfill these expectations?*

To explore these questions, we engaged 20 woman-identifying stakeholders in articulating ideal interactions with risk detection AI in dating apps. We chose dating apps as the context of study because they are well known facilitators of online-to-offline harm - especially sexual violence - with victims being disproportionately women [3] (in some samples all victims of online dating SV have been women [102])). Furthermore, AI has long played a key role in the online dating user experience through matching algorithms, yet implementation of AI for safety is still in nascent stages despite the literature having criticized the relative absence of safety-oriented features in dating apps [6, 37, 114, 143]. To explore RQ1, women in our study described ideal scenarios of using risk detection AI through 1) retrospective discussion of past experiences with risk in online dating to identify challenges and opportunities for risk detection AI in the user experience, and 2) prospective ideation of human interaction with risk detection AI aided by visual scenarios of a persona being informed of online-to-offline risk associated with various user profiles. To explore RQ2, women engaged in a participatory model building exercise [76–79] to articulate anticipated data sources, model features, and decision rules for a directly explainable risk detection model that would operate according to their envisioned human-AI interactions.

The findings show that women envision interacting with risk detection AI to augment rather than replace their existing, manually performed strategies for risk detection by expediting the collection and sense making of information they deem relevant to risk. Suggested data sources and features in the risk detection models they created were often subjective and admittedly fallible indicators of risk that held deeply personal significance based on their prior experiences of harm, such as a user's religion or education level. This discovery holds important implications for future research on user-centered design of risk detection AI that are unpacked in the paper. One is that **women in our study do not necessarily expect or want risk to be**

detected according to objective or statistically sound predictors of harm. Instead, they want risk detection AI to follow and amplify their own subjective (and biased) indicators of risk. This brings into question the practical utility of existing risk detection AI models that pursue a singular ground truth for risk. This also poses opportunity for research and development of user-personalized risk detection AI models and interactions. Yet related to this is the need for dedicated consideration of the ethical implications of subjective risk assessment AI, which may amplify harmful biases against already-marginalized groups.

The rest of the paper is structured as follows. First, we review prior literature related to online-to-offline harm and risk detection AI, followed by our method and findings, which explore the risk detection models produced by participants and conclusions about anticipated human-risk detection AI interactions. The paper concludes with limitations of the study and then discussion of the design and research implications of our findings for human-centered risk detection AI.

2 Background

In this section we first review online-to-offline harm and its prevalence in online dating, particularly against women. We then review technologies intended to address harm that manifests in the physical world to contextualize mounting attention from the research and public sectors on risk detection AI. We conclude by reviewing calls in the literature for involving stakeholders in the design of risk detection AI and methods for facilitating such involvement that serve as a backdrop for our own approach to involving women in the design of online-to-offline risk detection AI for dating apps.

2.1 Online-to-Offline Harm

Computer-mediated communication has enabled novel forms of interpersonal harm - especially against women and other marginalized groups (e.g., [84, 104, 107, 124, 126]). Such harm can range from cyberstalking [107] to abuse over social media [93, 126] and harassment in emerging social technologies like VR [52]. The literature has also reported on harm that spans across online and offline modalities such as revenge porn, referring to nonconsensual sharing of sexual acts filmed or photographed in the physical world [60, 61, 85].

A focus of our work is *online-to-offline harm*, which manifests in the physical world (e.g., sexual violence, bodily harm) due to facilitation by prior computer-mediated communication. This can include use of social media to lure victims into physical sexual abuse [116] or sex trafficking [100, 129], as well as doxing, during which home addresses are leaked through public social media posts, resulting in death threats and other hostile pranks like "swatting" - making false reports of heinous crimes to police to spark a Special Weapons and Tactics (SWAT) team response on an unsuspecting victim [49].

Dating apps have become a prominent facilitator of online-to-offline harm. Several studies have produced alarming evidence of sexual violence amongst online daters during face-to-face meetings (e.g., rape and unwanted touching of the body) [31, 34, 55, 102, 124, 133]. Approximately 10% of sexual assaults in samples from Australia [102] and the United States [124] were attributable to dating apps. Rates of online-dating-facilitated sexual assaults have increased through the years [4] - this increase is particularly notable amongst woman-identifying victims over the last five years [3]. Harm against users because of their gender and sexual identity [51, 81], as well as sexual risks such as HIV transmission [130, 131], are also persistent concerns. The physical harm facilitated by dating apps can be preceded by one or more of the many forms of online harm that also afflict online daters [127]. These include unsolicited nude imagery [127], financial scams [6, 119], bullying [21], racism [29], and sexual harassment [75].

The literature has been critical of a relative absence of safety-oriented design of dating apps [6, 114, 143], particularly for accommodating the needs of marginalized demographics [51, 97].

Some research portrays dating app design as playing an active role in the perpetuation of harm, such as by enabling perpetrators of sexual harm to find and manipulate victims into meeting face-to-face [124]. In addition, dating apps model sexual scripts that obfuscate perceptions of sexual agency [140] and promote sexual objectification [73, 82, 98]. They also scaffold sexual consent practices that predispose users to initiate physical sexual acts without asking permission because of assumed consent through indirect signals received in the dating app interface [140]. Prior work also argues that increasing rates of online-to-offline harm through dating apps [3, 4] are in part due to their broadened use for reasons beyond dating such as friendship and social activity partners [62, 96, 122, 123]. Leading dating apps have come to explicitly encourage such goals; for example, Bumble has designated sections in its interface for finding friends, business connections, and dating partners [62]. These varied purposes for app-use further obfuscate users' intentions for meeting face-to-face, thus exposing more people not only to online predators but inadvertent harm through misinterpretation of meeting goals - this can be especially risky with mismatches in sexual intentions [142].

2.2 Combating Online-to-Offline Harm With Risk Detection AI

Online-to-offline harm could potentially be stopped at the point of manifestation in the physical world. HCI literature has studied myriad wearable [90] and mobile [8, 108, 136] technologies intended to intervene in physical (sexual) violence. These devices can monitor women's safety through GPS [8], provide safe routes and emergency alerts about nearby assailants [22, 94, 108, 136], and alerts trusted contacts for assistance through use of a panic button [5, 68, 89, 103, 115]. The area of post-harm support has been given attention as well and could be valuable for victims of online-to-offline harm, such as with chat bots for providing support services to sexual violence survivors [85], and use of social media for support-seeking after sexual abuse [13] and reframing of sexual harassment experiences [45].

Another technology with strong potential for scalable mitigation of online-to-offline harm is *risk detection AI*. This refers to implementations of artificial intelligence in social technologies for identifying attempts at online harm or patterns of (online) behavior indicative of prospective harm in the physical world.

The types of risks that AI models have been developed to detect are wide-ranging. Regarding online-to-offline risks, prior work has focused on AI detection of online sexual child grooming using natural language processing of chat conversations [14] and detection of online ads and social media posts intended to coerce women and children into sex trafficking [129]. Detection and characterization of doxing posts on social media (leaking of home address and other personal information leading to offline harm) has also been studied [66, 113]. Conversely, risk detection AI has been developed to detect harm that traverses from offline to online, specifically skin detection AI to identify non-consensual and illegal sexual imagery [120]. Examples include the sharing of "upskirt images" that are taken in public without the consent or awareness of the victim [64], and detection of child porn [105] before it is posted online. Risk detection AI has also been a popular approach for identifying attempts at online harm, with cyberbullying being a primary focal point [71, 101, 106].

Within our focal context of dating apps, AI has played a foundational role in their design for years in the form of user matching algorithms. However, only recently has AI for risk detection become a more prominent focus with Bumble's open-source cyberflashing detection AI [121], and Tinder's detection of harassing message content [1]. Despite these efforts, there is a conspicuous void of *online-to-offline* risk detection AI in dating apps that can assist in mitigating physical harm that occurs during face-to-face meetings between users.

2.3 Stakeholder Involvement in Design of Risk Detection AI

Involvement of stakeholders in development of AI has been advocated across a variety of application domains [23, 35, 36, 117, 132, 134], and this holds true for risk detection AI as well. The literature has identified several concerns necessitating stakeholder involvement in risk detection AI.

To the first, risk detection AI is commonly trained on public social media data sets [53, 59, 70, 83, 118], which do not accurately represent real users and their interactions online [9] through omission of private messaging interactions where harassment, exploitation, and grooming often occur - in some cases data sets are comprised entirely of law enforcement personnel impersonating victims [100]. Another issue is the absence of stakeholder involvement in labeling data sets for training risk detection AI or in determining what qualifies as harm [71, 139]. Lastly, much of the risk detection AI research has focused on "computational aspects" [120] and capabilities such as accuracy, recall, precision of after-the-fact risk detection (see examples in [53, 59, 70, 83, 118, 135]). While these technical advancements are essential, there are persistent gaps in understanding how risk detection AI should intervene in user behavior or otherwise be implemented in social technologies to provide practical benefit to users [100, 120].

In responses to these concerns, HCI researchers have explored ways to involve users in the development of risk detection AI through participatory design methods [92] that position stakeholders as designers and decision makers rather than simply evaluators of researcher/practitioner-created designs. Two promising approaches have been through donation and annotation of private social media data to train risk detection AI models [9, 41, 99]. In practice data donation and annotation have been supported collectively in dedicated applications. Examples include an Instagram Data Donation application in which users donate their direct message (DM) conversations and label each conversation as safe or unsafe [99] and *MOSafely, Is that Sus?* [10], a dashboard for youth to donate their social media data, receive an overview of risks identified in their data, and provide feedback to the system on the accuracy of risk prediction. Prior work has also involved stakeholders in producing empirical insight [16, 56, 58, 65, 80, 137] for supporting users in uploading sensitive and personal data through these dedicated platforms.

Additional participatory methods could extend opportunities for inclusion, particularly in earlier conceptual stages of risk detection AI development to answer questions such as what data the AI should be trained on (and thus what data should be solicited for donation) and how user interaction with risk detection AI should be supported. Participatory AI design is still in fledgling stages [141] and carries unique challenges [18, 25]. Nevertheless, diverse approaches to stakeholder involvement in AI have already been showcased in the literature (e.g., [11, 74, 91, 110, 138]).

A method that has been applied to other participatory AI design contexts, but not yet risk detection AI specifically, is participatory building of explicit rule models. This entails stakeholders articulating specific features to be incorporated in an AI model along with decision rules for those features (see examples for food donation allocation models [78] and worker well-being models [79]). Participatory model building can be similarly beneficial to risk detection AI for dating apps because online daters - especially women and other marginalized groups - have long engaged in personal strategies for managing safety [19, 51, 87, 97] and assessing potential meeting partners [50] (called impression formation [21] or uncertainty reduction [54]). For example, users have reported avoiding profiles with pictures that obscure one's face [21, 75]. Some require video calls before meeting in-person to confirm identity [75], whereas others verify a meeting partner's information through search engines [54] or seek out third party information that may be more trustworthy than self-reported profile content [125]. Ultimately, participatory model building can enable online daters to leverage their personal experiences with risk management to inform design of risk detection AI and subsequent stages for stakeholder involvement such as data donation and annotation.

3 Method

We conducted an IRB-approved study with woman-identifying dating app users in the United States and Canada (n=20) to understand how end-users envision interacting with risk detection AI (**RQ1**) and how risk detection AI should be designed pursuant to those interactions (**RQ2**). The method involved interview sessions with individual stakeholders, blending elements of end-user elicitation and participatory design to reflect on past experiences with risk and prospectively design risk detection AI models operating according to envisioned scenarios of user interaction with it for risk assessment of meeting other dating app users face-to-face.

3.1 Participant Recruitment

We focused on recruiting woman-identifying users given the severely disproportionate victimization of women in online dating [3] and shared gender identity with the researchers moderating interview sessions (see rationale in the next section on precautions for participant comfort). Recruitment channels included online advertisements on various subreddit forums related to online dating (r/bumble, r/tinder, r/Zoosk, r/coffeemeetsbagel, r/Plentyoffish, r/hingeapp, r/Datingapps, r/OkCupid), email lists associated with our university student body, word-of-mouth campaigns amongst woman-identifying members of university clubs and local community organizations catering to women, and snowball sampling. Inclusion criteria necessitated participants identify as women and have prior online dating experience. Participants were compensated with a \$30 gift card.

Ages of participants ranged from 18 to 41. Participants identified as black (6), white (6), Asian (4), mixed race (2), Native American (1), and middle eastern (1). Participants resided in 11 different states around the United States, and one resided in Canada. While we did not ask if participants were cis- or transgender women, two voluntarily disclosed as transgender. The most popular dating app used by participants was Tinder (n=9), followed by Bumble (4), OkCupid (3), Hinge (3), Facebook Dating (1), Coffee Meets Bagel (1), Muzmatch (now called Muzz) (1), Match.com (1), Christian Mingle (1), eharmony (1), and Zoosk (1). Some participants also reported using other social media apps for discovering dating partners, including Instagram (4), Meetup (2), Facebook (1), Snapchat (1), and Patio (1).

All but 6 participants reported face-to face encounters with dating app users that were risky, unsafe, or harmful. See table 10 in the appendix for demographic details of all participants and their corresponding experiences with risk and harm.

3.2 Precautions for Participant Comfort

Consultations with sexual violence researchers and a sexual assault nurse examiner (SANE) informed the method for this research, who collectively have several years of experience interacting with victims and perpetrators of sexual harm in research and/or clinical contexts. Additionally, our SANE consultant referenced the trauma-informed approach (TIA) as the basis for her recommendations, which is a well-established framework for supporting trauma survivors [63] (see other ways TIA has been applied in HCI research in [32, 111]).

Recruitment materials clearly stated that the study pertained to online dating harm, allowing prospective participants to make informed decisions about potential retraumatization. Private, individual interviews gave participants control over who witnessed their disclosures of risk experiences. A woman-identifying researcher led all sessions to build trust and familiarity, with another woman-identifying researcher providing technical support and taking notes which ensured the lead moderator could be attentive to the participant. Researchers emphasized participants' rights to

skip questions and end sessions early without it affecting their financial compensation. Additional participant care structures are detailed in the data collection section.

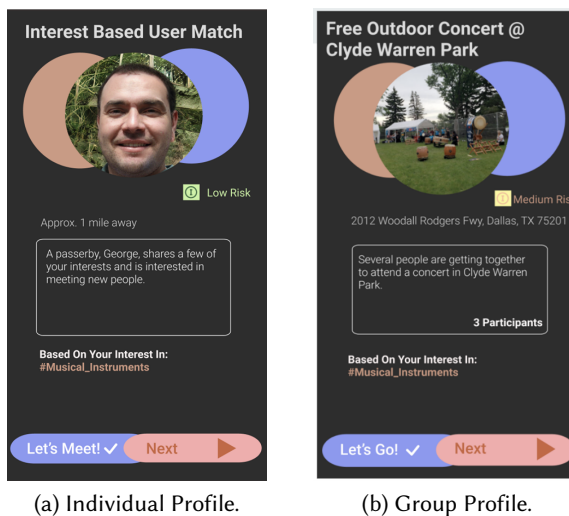
3.3 Data Collection

The interview sessions were conducted online over Zoom, ranging from 1 hour and 51 minutes to 2 hours and 19 minutes. **Activity 1 - past experiences with risk:** First, participants reflected on their personal experiences with dating apps, followed by encouragement to expand on a particular experience that they found risky, unsafe, or harmful. This component of the protocol was intended to recognize participants as experts of their own safety, reinforcing their authority—not the researchers’—on how risk is conceptualized and should be assessed by AI or any other entity. This was essential given sexual violence research [69] and practitioner advice highlighting systemic issues in loss of agency over risk/harm experiences by women when disclosing to authority figures.

Activity 2 - ideation of human interaction with online-to-offline risk detection AI (RQ1): Participants then ideated human-AI interactions within a hypothetical dating app to foresee risk of harm in meeting a given user face-to-face. Participants then reflected on how a persona user named "Anna" should interact with risk detection AI during profile discovery and messaging with other users in the dating app to foresee risk of harm associated with meeting face-to-face. Grounding ideation with a persona was at the recommendation of our method consultants to allow dissociation with prior (and potentially traumatic) experiences. Priming materials included a presentation on AI in dating apps and prototypical social matching apps from prior HCI research (e.g., [86]), as well as the conceptual workings of risk detection AI such as labeling datasets of risk. However, specific examples of existing risk detection AI were withheld to avoid biasing their understanding of how it could or should operate.

Activity 3 - risk detection model building (RQ2): Following participatory model building studies in other contexts [76–79], participants built directly explainable risk detection AI models. This involved specifying model features/variables, decision rules (i.e., how each feature’s potential

Fig. 1. Model building was motivated by profiles conveying a hypothetical output of risk detection AI (e.g., "medium risk"). Profiles were fabricated according to each participant’s social interests.



state influences risk determinations), and envisioned data/input sources that inform the model features. An end-user elicitation technique [7] contextualized this activity. First, participants were presented with a series of 3-4 profiles that the persona Anna would discover in the dating app, each featuring a "risk level" indicator (low, medium, or high) representing the output of a hypothetical online-to-offline risk detection AI (see Figures 2a, 2b). While the risk indicators were randomly assigned, the social opportunities depicted in each profile were custom-made for each participant according to screening survey responses about social goals, interests, and activities that drive app use. Profiles represented a mix of individual people and group-based activities, informed by mounting evidence that dating apps are being used [20, 96, 122, 123] and designed [48, 62, 122, 123, 142] for multifaceted user goals. By having participants reflect on multiple types of social opportunities their resultant risk detection AI models would thus be generalizable to multiple contexts of online-to-offline risk.

Participants created a risk detection AI model reflecting their perception of how the AI could, or should, have computed the respective risk level for each presented profile. A model template supported listing of specific model features, possible states of each feature, and weights associated with each state. A profile's overall risk level would be determined by a summation of the numerical weights for every feature in the model. See Table 11 for an example of the empty model template, and Table 12 for an example of how a feature for "criminal record of the other person" factored into P4's model, both in the Appendix. See the supplementary materials for P4's completed risk detection model template.

These risk detection AI models were constructed collaboratively with researchers using the template, enabled through a shared screen. After a tutorial of the template and its key elements the participant would engage in open-ended verbal speculation on the model's features, which the researchers recorded in the template for the participant to further edit and elaborate on. Typically, participants began suggesting specific features, feature states, and weights on their own as they became more accustomed to the terminology and template. The following quote exemplifies a collaborative exchange between P13 and the researcher about a feature for gender of dating app users informing risk:

Interviewer: *About gender, you mentioned for mixed [men and women mixed in a group] and women [only] you will feel better than if all are men in a group? How can we evaluate that? If you want to define that in another way, feel free to.*

P13: *I think that that's a good way to define it. I feel like if they were men, then it [the weight for that feature state] would be probably an 8 [out of 10], if it was 100% women or women presenting people, I'd say it'd be about a 1 or 2. And then if it was a mixed group then that's about a 3 or 4.*

Participants populated a single model template with features applicable to all their example profiles. They then computed a separate total risk score for each profile by imagining the relevance and state of each feature per profile and adding together the numerical scores for each feature. If the total risk score felt incompatible with the risk level indicated on a sample profile (e.g., if the total score for a "low risk" profile felt too high to them) the weights of some feature states were modified to bring the total score down. The exercise concluded with discussions of the personal data that participants would donate and/or expect other users to donate to enable proper functioning of their risk detection model.

3.4 Data Analysis

Artifacts produced through the study included 1) risk detection model documents (see supplementary materials for each participant's model), and 2) audio recordings and auto-produced transcripts

of the interviews. We subjected the data to reflexive thematic analysis (RTA) [26] to answer our research questions. RTA was chosen for its suitability to capture both semantic and latent meaning in the data [27] and its flexibility with data sources that could be subjected to analysis [24, 28], which were prerequisites for our analysis given the coupling of semantic data from risk detection models and the latent motivations behind anticipated human-AI interactions from interviews.

RTA involves six steps [27]: 1) familiarization with the data; 2) coding; 3) generating initial themes; 4) developing and reviewing themes; 5) refining, defining, and naming themes; and 6) writing up results. Explication of, and reflection on researcher positionality to data is crucial throughout this process [30]. Our analysis was conducted by two researchers: a woman-identifying researcher (steps 1-6) with experience in personal risk of harm and publishing on computer-mediated sexual violence, and a heterosexual man-identifying researcher (steps 4-6) with professional experience interacting with victims and perpetrators of sexual violence in research contexts. Reflexivity exercises acknowledging the influence of these experiences on data analysis are noted with their respective steps below.

Familiarization of the data (step 1) was performed first by consolidating participants' risk detection models into a single table for easier review and identification of similarities across models. See Table 13 and a description of its consolidation process in the Appendix. We continued data familiarization by proofreading and revising auto-generated transcripts from the interviews, along with personal note-taking to explicate positionality to the data (e.g., personal experiences that resonate with those of participants).

The coding stage (step 2) involved line-by-line coding of the interview transcripts in the qualitative analysis software Dedoose. The consolidated risk detection model table was not included in this step; see step 6 for how the model table was incorporated in analysis. Coding of the interviews was primarily semantic initially, as exemplified with codes for types of harms and specific features for risk detection models spoken verbatim by participants. As coding progressed into initial theme generation (step 3), latent codes were added that captured the interconnection between personal experiences with risk and envisioned uses of risk detection AI. Miro [88] (an online whiteboard tool) was used for refinement of themes (steps 4-5) with the second researcher through multiple synchronous meetings. The researchers' personal and professional experiences with risk were foregrounded in discussions around coding hierarchies and re-organizations, often due to differing perspectives on themes (e.g., whether emotional harm warranted its own category or could be situated alongside other non-physical harms).

For the write-up of results (step 6), we transformed the visual thematic map from Miro into paragraphs, during which the consolidated risk detection model table was formally introduced to the thematic mapping. This often involved production of new variants of the consolidated table to enrich themes from the interview transcript analysis. For example, a theme from the transcripts was that participants' risk detection AI models sought to detect the user's ability to seek help during a face-to-face date. Accordingly, we created a new table that listed specific model features associated with that type of detection, and another to demonstrate their popularity relative to other features. Because these tables were incorporated into the thematic mapping upon creation we report them in conjunction with participant quotes in the Findings section.

4 Findings

Participants consistently indicated they did not want risk detection AI to replace women's manual attempts at assessing risk of harm in online dating. Rather, the intended role of risk detection AI was to augment their existing strategies for self-protection. **Augmentation of women's personal safety strategies manifested in two phases of human-AI interaction: a) preparing the AI to detect risk according to the user's subjective indicators of safety, and b) evaluating and**

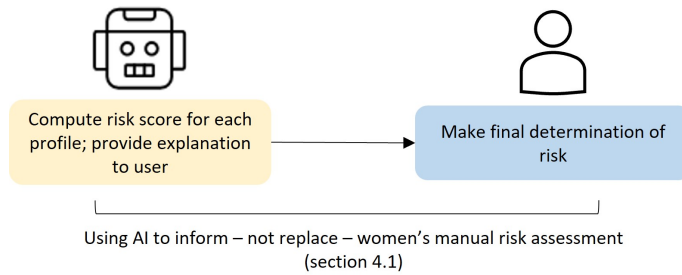


Fig. 2. The anticipated role of risk detection AI in risk assessment (see section 4.1).

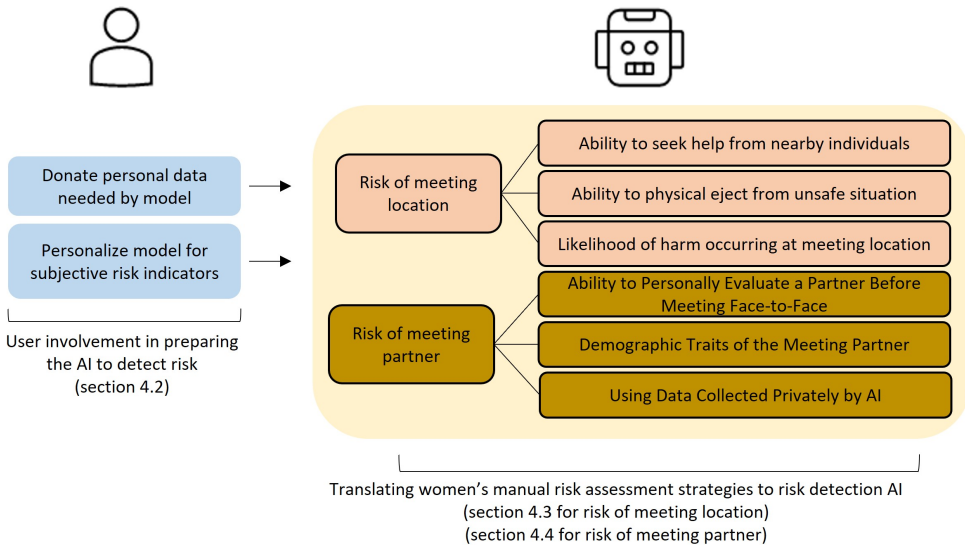


Fig. 3. The envisioned functioning of risk detection AI models (see sections 4.2, 4.3, 4.4)

synthesizing the AI’s risk computation into one’s personal assessment of risk. *Preparing the AI to detect risk* was imagined as a crowdsourced data donation exercise in which all dating app users would provide personal data about themselves used by the risk detection AI to compute risk scores for every user (including themselves). This preparation stage would also involve personalization of the risk detection AI model to the individual user through modification, addition, and deletion of features in the model to reflect the user’s subjective risk detection strategies (e.g., adding religion of a meeting partner as a subjective indicator of risk). *Evaluation and synthesis of the AI’s risk detection into one’s personal assessment of risk* would occur at the point of discovering and interacting with a potential meeting partner through the dating app. This would involve reviewing explanations of how the AI computed its risk score, along with disclosure of data sources, to help a user determine if and how to incorporate the AI’s risk assessment into their personal determination of risk for a given social opportunity.

Participants constructed their risk detection models to largely mimic their manual strategies for assessing risk of harm. Risks during face-to-face encounters that participants sought to detect fell into two categories: physical and non-physical harm. Physical harm was typically exemplified as

sexual in nature, whereas participants identifying as transgender were also concerned with physical violence on the basis of their gender identity. Non-physical harm constitutes harm unrelated to physical violence, such as emotional harm stemming from hurtful comments. Manual strategies for assessing risk focused on two data sources: information about the face-to-face meeting location that may affect one's ability to respond to risk and harm, and information about one's meeting partner(s) that may imply their likelihood of inflicting harm. **Participants' risk detection AI models were intended to improve on and support women's manual strategies for assessing risk through faster, more consistent, and more granular collection of risk-relevant data.** As a result, proposed features for risk detection AI models fell into six more precise categories. For meeting location these included 1) *ability to seek help from nearby individuals*, 2) *ability to physically eject from an unsafe situation*, and 3) *likelihood of harm occurring at the meeting location* (regardless if committed by one's meeting partner or other bystander). For meeting partner these included 4) *the ability to personally evaluate a partner before meeting them face-to-face*, 5) *risk associated with demographic traits of the meeting partner*, and 6) *risk identified in data collected privately by AI*.

Figure 2 depicts the envisioned functioning of risk detection AI models and Figure 3 depicts how users would interact with said models. The remainder of the Findings section starts with how users would incorporate the output of risk detection AI into their own risk determinations (section 4.1), followed by users' actions to enable risk detection AI models to augment their manual safety strategies as intended (section 4.2), and concluding with an analysis of participants' risk detection AI models themselves, divided into risk associated with face-to-face meeting locations (section 4.3) and risk associated with the face-to-face meeting partner (section 4.4).

4.1 Using AI to Inform—Not Replace—Women's Manual Risk Detection Efforts

"Regardless of what my risk evaluation [AI says], I would still make the decisions on a case by case basis. Nothing is a total stop or total go, whatever you have from the application is just to assist your decision." (P6)

Participants consistently stressed that users, not the AI, would make a final determination about risk associated with a given face-to-face encounter. While there was resistance to the prospect of risk detection AI as a risk *authority*, participants were quite receptive to its inclusion in the online dating user experience as an *assistant* to their manually-performed strategies for detecting risk as alluded to in P6's quote. Key to this assistant role was the expectation that the risk detection AI would offer an explanation for how it computed its risk score for each dating app profile. This was mostly for context that would help users assess if and how to incorporate the AI's risk detection into their personal determination of risk. Another reason several participants wanted an explanation was to validate that the model was identifying risk in ways that adhere to their personal strategies.

"Like if it was going to have low risk, medium risk, high risk, I would like descriptions of why, you know, that would be helpful, as opposed to just the algorithm making a decision for me. And it's gathering, I have given it information [about how I determine risk], such as what I'm comfortable with." (P11)

Some examples of explanations included identifying the data that informed the model's understanding of particularly subjective indicators of risk. P14 gave an example in which the AI may determine a user to be high risk due to being an alcoholic, which could be a subjective conclusion depending on the data used to deduce alcoholism. Explanations of the model's output would provide users the opportunity to determine if they agree with how the AI came to its conclusion of risk. P14, referencing a profile with a high risk indicator: *"What information could the app be collecting to make this especially high risk? I don't know. I mean, he's drinking a beer in his profile picture. Was he an alcoholic or something? I don't know. It's important, but then it also makes me nervous if [...] like we have deduced they're an alcoholic [without me seeing why]."*

Other needs for explanation stemmed from potential disagreement between users and the AI around the significance of particular data points. A few participants imagined scenarios in which a potential meeting partner has a record of criminal offenses of varying degrees, some of which may not be meaningful or relevant to assessing risk from the user's point of view. P7 provided an example in which a potential meeting partner may have been cited for jaywalking, which they did not personally consider relevant to risk. P14 similarly considered misdemeanor crimes as less relevant to risk. In P7's words: *"I would like to know why because what if they're a minor [offender] or if they are a person with a minor offense on the record, for something really petty such as jaywalking, I'm not gonna care about that. But the system might care and say, oh, let's medium [risk] this guy."*

Beyond explanations of the risk detection AI's decision-making, a few participants requested that the AI be able to share the data that informed its risk determination, so that the user could incorporate such information into their own risk detection processes. As P12 put it: *"I would actually like some information. In general, I would actually like the application to provide me with the information [used to determine] low risk, high risk, medium risk."* Requests for the data underlying the AI's risk decisions stemmed from consistent struggles that participants had in executing their manual risk detection strategies in the past due to limited availability of information in profiles and messaging interactions that they considered informative to risk.

Participants were cognizant, however, of privacy concerns if risk detection AI were to make publicly accessible *all* data about other users that factor into risk decisions. They reflected on the likely discomfort they would feel if their own data was fully reviewable by other users when assessing risk that they themselves may pose. In P14's case, they considered but quickly changed their mind about having full access to the information used by the AI: *"It's a two way road, they have access to you, and you don't know who they are. And I feel like it would just be safer all around [for users] to not have access at all."*

4.2 Preparing AI to Detect Risk According to Women's Manual Safety Strategies

The risk detection models created by participants were intended to improve on and support their manual strategies on dating apps for assessing risk associated with 1) anticipated face-to-face meeting locations and 2) anticipated face-to-face meeting partners. The improvements that risk detection AI would make to their manual safety strategies stemmed from an improved availability of risk-relevant data. Most instances of manually-conducted risk assessment reported by participants we describe as *data-poor* in the sense that their personal risk detection capabilities were dependent on often-scarce and inconsistently available data from user profiles and messaging interactions, along with first-hand familiarity of meeting locations. The proposed risk detection AI models were comparatively *data-rich* in the sense that they sought to provide uniformity and predictability to risk detection through more consistent access to data indicative of risk.

Participants recognized that the feasibility of their proposed risk detection models would be dependent on availability of data related to the features of their models. As such, the brunt of imagined user involvement for preparing risk detection AI was expected to go towards helping the AI collect the many data points necessitated by participants' models. Three user activities to aid in this data collection were mentioned most frequently: private profile pages only viewable by risk detection AI, social media account linking, and user reports about transient risk-relevant data.

4.2.1 Private Profiles for AI Model Training and Customization. Some participants imagined an additional phase of account setup; in essence, creating a public profile for other users to see and a second private profile of personal data only accessible by the risk detection AI. Examples of personal data that users may provide to their private, AI-only profiles included their religion, marital status, education level, mental health problems, political affiliation, gender identity, and so on. Some

participants expected there to be a stark disparity in information users would be willing to provide to private profiles for AI model training compared to public profiles viewable by other people. Per P5: *"There are some things about him that the app knows that I do not. If the app can provide [that data to my personalized risk detection model], then that'd be really nice."* Importantly, P5's quote is indicative of how many participants viewed the practicality of personalized risk detection models as a crowdsourced responsibility. All users would need to disclose personal information about themselves - regardless of whether that information is applicable to their own risk detection model - to enable the functioning of other users' personalized models for whom such information is necessary.

Several participants expected these private profiles to also serve as a terminal for collaboration between the risk detection AI and the user to tailor identification of risk to a user's personal, subjective risk detection strategies. This would prepare the AI to detect risk in similar ways as the user would, albeit with more efficiency. Customization was typically imagined as a visual modification of the model that each participant built through modifying weights of existing features as well as adding or removing features that the user deemed (in)applicable to their personal assessments of risk. Participants expected this model personalization to be an ongoing process; something that users recurrently update as they continue their dating app-use and meet new people.

4.2.2 Social Media Account Linkage. Some participants envisioned their risk detection AI having access to users' social media profiles for collecting information about personality and demographic traits deemed risk-relevant like political and religious views. Yet participants also noted privacy concerns and general discomfort with making personal social media accounts available for risk detection. P14 mentioned a past situation with a stalker tracking her public social media accounts and considered sharing social media access with all users to be *"uncomfortable."* Nonetheless, P14 alongside other participants indicated willingness to share their own social media accounts privately with the AI so that it can *"do its job."* As they elaborated: *"All of my information, I'd be happy to share with the application as long as it's not shared with other people. So that the application can still do its job to assess, like, hey, these are what I see as risks."* When it comes to having other users share their social media access to the AI, participants also commonly emphasized the need for transparency and *"consent"* from users: *"I think consent is really important. Like if users know exactly like how their information will be used, and like what exactly is being collected, that's, that's super important"* (P14).

4.2.3 User Reports on Transient Risk-Relevant Data. Participants often noted third-party data sources that they expected their risk detection AI to have access to for computing risk, such as Google and Yelp reviews to amass understanding of location type and the typical presence of bystanders. However, they also noted that these third-party sources can be unreliable, outdated, or incomplete. In these instances several participants suggested that risk detection AI could actively solicit *"reports"* from users to inform risk-relevant data points. One example was the collection of cell phone/wifi service reports from other users/visitors at a popular face-to-face meeting location, which may impact one's ability to call trusted contacts for help if they choose that location for a date. P14 described her enthusiasm for the idea: *"I would love it. If they're like, hey, if you do go to this location, we've collected data, or we've received multiple reports that cell service drops in this area."*

Another, less common, idea for user reports involved submitting reviews of other users after meeting them face-to-face, which could inform a safety *"reputation"* score for each user. For instance, if a user felt unsafe during a face-to-face encounter, they would submit a report to the risk detection AI that would increase the risk score on that partner's profile when discovered by others. P1 reflected on the pros and cons of user-submitted reviews: *"So maybe what it will be is something like people that have [met] can leave a comment, like, 'oh, they're really fun' and submit that. [Although] I feel like a rating system, even if it is as simple as a thumbs up or a thumbs down, that might be kind of hard"*

to keep equitable." P1's concern of "equitable" ratings was shared by others who worried of abuse and intentional bias, and ultimately this data preparation activity was not strongly recommended even by those who initially thought of it.

4.3 Designing AI to Detect Risk Through Meeting Location

Information about an eventual meeting location was an important consideration in participants' manual risk assessment strategies because of the role location can play in abetting or preventing harm. Likewise, location-specific features were common in participants' risk detection AI models. Our analysis organized such features into three categories of risk associated with face-to-face meeting locations: 1) *the ability to seek help during a face-to-face meeting*, 2) *the ability to eject physically from an unsafe situation/location*, and 3) *the perceived likelihood of violent acts occurring at the location* (independent of their meeting partner).

4.3.1 Detecting Risk Through Ability to Seek Help During a Face-to-Face Encounter. An important contributor to online-to-offline risk for participants was the ability to seek help should a face-to-face encounter become perceptually unsafe. In their personal strategies, the entities that participants described getting help from included friends and family present at the group event or through phone calls or text messages, as well as individuals around their physical location such as professional security staff affiliated with the meeting location and other patrons ("bystanders") - especially families or groups of people that could provide protection in numbers. The presence of families at the meeting location was valued not only for the assistance they could provide, but as a possible deterrent for harm ever occurring. As P2 described: "*The most dead giveaway when it comes to safety is regardless of how ill-intentioned a person might be it's very difficult to expect that anyone is going to do anything that stupid, crazy or awful in front of children.*"

Table 1. Features indicative of ability to seek help during a face-to-face encounter.

Features	Contribution to risk assessment
Bystander presence	Strangers at the meeting location could provide immediate intervention during an unsafe encounter
Location population density Location publicness	Densely populated areas could be an indirect indicator of bystander availability
Meeting start time, end time Day of week for meeting	Availability of bystanders varies with time of day at some locations
WiFi Phone reception	Phone connectivity enables the user to call friends, family, and emergency services for help
<i>Familiarity between meeting partners*</i>	Attendees of a group event who are already friends may be less willing to intervene if one of their friends is posing risk of harm
<i>Presence of friends</i>	Friends in a group-based social activity are more likely to provide assistance than strangers
No. of meeting partners	Presence of other people for a group-based social activity could deter an individual from causing harm
Location familiarity	Users may be more willing to seek help from familiar staff at a meeting location they have been to before
Security and police proximity <i>Presence of group activity host*</i>	People in authority may be more likely to provide assistance

*Italicized features exclusively pertain to group events

Participants’ AI models had 14 features related to ability to seek help. These pertained to both the presence of bystanders and friends (or at least familiar attendees) at their location as well as the capacity to contact trusted friends and family with their phones. However, analysis of their models produced a clear prioritization of availability of nearby strangers over capacity to contact trusted family and friends who are not physically co-located with them. This was rooted in a desire for “immediate” assistance should harm occur, which trusted contacts would be unable to provide. Relatedly, several participants considered group-based social opportunities to be inherently safer than meeting an online dater alone because of the immediate assistance that fellow activity partners could provide. As P17 described, attendees of a group-based activity are in “the same boat” and can enact a mutual safety structure.

Table 2. Visualization of model features indicative of ability to seek help during a face-to-face encounter. Most features relate to the availability of bystanders for immediate assistance. The features are sorted according to the number of participants that had the feature in their model. The numbers within each cell reflect the maximum weight that the participant gave to the feature, normalized across a 1-6 scale with 1 reflecting the least risk and 6 reflecting the most risk.

Participant	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P12 (g ^{**})	P13	P14	P15	P16	P17	P18	P19	P20
No. of meeting partners	5.4	5.4	6	4.5	6	6		4.3	3.8	2.3		4.9	4.9	3.8	6	4	6	5.4	3.8	4.8	6
Location familiarity	5.4	4.3	6			6	4	4.5		3.5	5	5.4	4.9	6	4.8	5	6	3.8	4.9	4.8	3.5
Bystander presence	6	5.4	6	6	6			3.8	4.9		6	4.3	4.3		5.4		6	4.9	5.4	4.8	6
Meeting time	6	4.9	4.6	5.8	5.2	6			3.8			4.9	4.9		3.5	6	6	5.4	6	6	6
WiFi/phone reception				6			6	5.5	6	6		6	6	6	6	6	6				6
Security/police proximity				5.5			4	3.8				5.4	4.3	5.2		3	6	3.8			
Location publicness			6			6			5.4						6	6		4.3			
<i>Presence of friends*</i>			4.6							3.5	4.5							4.9			6
<i>Familiarity b/w meeting partners*</i>				5.4			5											5.5			
Location type		3.8														6		3.2			
Day of week for meeting			6		4.3																
Location population density										3.5					3.5						
Meeting end time																	5.3				
<i>Presence of activity host*</i>											6										

*Italicized features exclusively pertain to group events. **P12 made a separate model for group events.

In analyzing the features in participants' models for ability to seek help (Table 5), the most common and most heavily weighted features were directly or indirectly related to immediate bystander intervention. The most blatant examples include number of meeting partners (referring to the number of people from the dating app who are participating in the respective social activity) as well as presence of bystanders who are unaffiliated with the dating app-arranged social activity but would still be available for assistance. Features that indirectly pertain to bystander intervention include familiarity with the meeting location because that would be indicative of the user's rapport with staff at the location and comfort level with approaching them for help. Another example is meeting time because there may be more bystanders present at "peak" times (e.g., meeting at a bar in the evening as opposed to the afternoon). Although we should note that one participant, P16, considered locations or group activities that were *too* crowded, like popular bars or night clubs, to reduce capacity for seeking help because pleas for assistance may go overlooked or unheard amongst a large crowd preoccupied with other activities. As P16 summed up her position: *"I think having too large [of a group size], that's just as risky as having too small."*

The prioritization and frequency of features related to co-located bystanders does not mean that trusted friends and family members were completely devalued in assessing one's ability to seek help. Wifi/phone reception was still a commonly suggested model feature as a reliable 'last resort' for help, as P20 encapsulated: *"I want to add cell phone connectivity [to my personal risk assessment model], that's very important, because my phone must have enough connection to be able to call for help if needed. I wouldn't want to be stuck."* Some participants referenced specific contexts where Wifi and phone reception would be disproportionately important, such as during face-to-face meetings in "remote" or rural areas that may not typically have many bystanders available for intervention. In these cases, being able to reach a friend or family member, or emergency service, may be the only way to seek help, even if that help cannot feasibly intervene immediately. P2 described the importance of this model feature during such encounters: *"I'm a city girl, I feel safer in the city, because my brain sort of already assesses those risks. You know, I have my phone, I know where to go. [...] If I was in a remote area, I would feel very uncomfortable. For me, if there's no cell reception, that's very remote."*

4.3.2 Detecting Risk Through Ability to Physically Eject From Unsafe Situations. While not as popular as other model feature categories, some participants also assessed risk by their perceived capacity to escape an unsafe situation, independent of the availability of bystanders to help them. The most common of these features was the proximity of the meeting location to one's home, which participants viewed as a safe haven. This sense of safety within one's home was partly due to the ability to physically barricade oneself from an unsafe person. Others mentioned the presence of trusted individuals at their home, such as family members and roommates, who could be relied on for safety. Reaching this safe haven, however, may not be an easy task in a dangerous situation, particularly if participants do not have a means of private transport (cars) or they are parked far away from the meeting location. As such, a few participants also included a feature for proximity to transportation.

It should be noted that the limited popularity of features around proximity to home and proximity to vehicle may be influenced by participants' geographic location and the fluctuating relevance or definition of proximity based on their locale. Similarly, while closer proximity to one's means of transportation was clearly considered safer than a farther proximity, the nature of "close" varied. For example, P8 lives in an urban area and therefore described a feature for proximity to public transit rather than a vehicle. P8's description of her "proximity to public transit" feature acknowledged that even the most convenient (and therefore safest) public transit access may be a few minutes away. By comparison, participants with vehicles expected to ideally park their vehicle in a parking lot

Table 3. Features indicative of ability to physically eject from unsafe situations

Features	Contribution to risk assessment
Proximity to home	Enables users to distance themselves from dangerous meeting partner(s)
Proximity to vehicle or public transport	
Easy to enter/leave meeting location	

right next to their meeting location - a multi-minute proximity to their vehicle would be considered very high risk. This difference in what qualifies as a safe proximity for vehicles versus public transit emphasizes that the nature of risk may be highly contingent on the mode of transportation and the geographic location of the user.

Relatedly, a third feature introduced by two participants was the ability to easily and discretely leave a meeting location should it become unsafe. P11 preferred public, open spaces for face-to-face meetings because of the relative ease of movement: *“I can leave when I want to, it’s not like I’m going to be locked in there. It’s just the ability to freely come and go, I guess is one of the things that can make a person feel safe.”* Examples of such locations included parks, libraries, and museums. An example of a location suggested by participants that *would not* be easy to leave included an isolated hiking trail because the terrain may prevent them from quickly leaving the location. On the contrary, overly packed bars and nightlife spots were also considered hard to leave because an excessive amount of people may prevent easy movement to an exit. P10 described the importance of this feature in creating a sense of safety through being able to leave a location whenever they like: *“When it is the sort of open, easy to access [meeting location] and I have control over when I’m arriving and when I’m leaving, I feel low or no risk.”*

Table 4. Visualization of model features indicative of ability to physically eject from an unsafe situation. The features demonstrate one’s home as a safe haven, and capacity to easily reach home as an element of perceived risk. The features are sorted according to the number of participants that had the feature in their model. The numbers within each cell reflect the maximum weight that the participant gave to the feature, normalized across a 1-6 scale with 1 reflecting the least risk and 6 reflecting the most risk.

Participant	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P12	P13	P14	P15	P16	P17	P18	P19	P20
													(g*)								
Proximity to home	4.9			4.8		6		4		2.9				4.6							
Easy to enter/leave location						6														4.8	6
Proximity to vehicle/transport						6	6	5													

*P12 made a separate model specifically pertaining to group based events.

Lastly, P10 emphasized an interconnection between model features for ease of leaving a location and bystander intervention (from the previous subsection on ability to seek help). Knowledge of being observed by bystanders could moderate the behavior of their meeting partner, particularly any retaliatory actions if they were to notice an attempt to leave the meeting location early. As she described it, being observed by others could put a meeting partner on their *“best behavior”* - in

effect, bystanders would be socially enforcing the ease with which P10 would be able to leave a location quickly. This is manifested in P10's model through a max risk weighting for both features.

4.3.3 Detecting Risk Through Likelihood of Harm Occurring at the Meeting Location. While bystanders were described in a positive light in the "ability to seek help" category of features, participants also discussed situations in which bystanders - and the activities they engage in - could increase perceived risk of harm. This was most commonly encapsulated in a feature that some participants called the "reputation" of a meeting location. This reputation would be informed by typical activities occurring at the location as well as the surrounding area. Participants' descriptions of reputation were often vague however. When describing how location reputation factors into their manually-conducted strategies for safety they typically made reference to socially learned understandings of which towns and areas are safe and which are associated with crime - or what is typically called 'gut instincts'. They also inferred reputation through the nature of the business establishment. For example, they ascribed riskier reputations to bars and nightclubs because of sexually charged behavior and use of alcohol and other drugs that they associated with these businesses.

Table 5. Features indicative of harm occurring at the meeting location

Features	Contribution to risk assessment
Location reputation	Socially learned perceptions of which geographic areas and businesses are associated with crime or reckless behavior
Location cleanliness	Low maintenance was considered a signal that safety may not be prioritized at the location
Presence of alcohol Recreational drug use	The presence of alcohol and other drugs was perceived as conducive to erratic behavior that could inadvertently lead to harm

When translating their assessment of location reputation to risk detection AI models, most participants kept a general feature for reputation that encapsulated socially learned understandings of towns and businesses which they would teach the AI through their private AI-only profiles (see section 4.2). Some also supplemented this with more granular features. The presence of alcohol was the most common of these, followed by recreational drug-use, because they associated these activities with reduced inhibitions that could lead to reckless behavior. Here participants made a point to distinguish risk of intentional attempts at harm from risk of inadvertent injury due to their vicinity. P9 also had a feature for location cleanliness, which they extrapolated as an indirect signal that the location and surrounding area lack neighborhood resources for safety such as security, a police station, or cell phone reception, drawing an interconnection to the "ability to seek help" category of features.

Table 6. Visualization of model features indicative of the likelihood of harm occurring at the meeting location. The features reflect the socially learned reputation of geographic areas and businesses, along with associations made between activities like alcohol consumption with risk of inadvertent harm. The features are sorted according to the number of participants that had the feature in their model. The numbers within each cell reflect the maximum weight that the participant gave to the feature, normalized across a 1-6 scale with 1 reflecting the least risk and 6 reflecting the most risk.

Participant	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P12	P13	P14	P15	P16	P17	P18	P19	P20
Location reputation		5.4		5	5.7				4.9	5.4	5.8	6	6		3.5		6		4.9	6	5

Presence of alcohol			5.8	6								3.5	6				3.5
Recreational drug use						5								3.5			
Location cleanliness							4.9										

*P12 made a separate model specifically pertaining to group based events.

4.4 Designing AI to Detect Risk Associated With the Meeting Partner

In addition to location-based indicators of risk, much of our participants’ manually conducted safety strategies put an emphasis on the perceived *"likelihood"* that their meeting partner may harm them. Indicators of perceived risk associated with meeting partners varied drastically from participant to participant. These signals of potential harm were often subjective and deeply personal, informed by participants’ past experiences of harm and feelings of danger through dating and other social interactions. For example, two participants identified education level as a reliable indicator of a partner’s propensity for harm, whereas others thought a partner’s religion was applicable to risk, or political affiliation and so on.

As a consequence of these highly varied personal strategies for assessing a meeting partner’s likelihood of causing harm, the proposed risk detection AI models exhibited 24 different features related to the meeting partner, with most of these features being mentioned by only a few participants or less. This is a stark difference with location-based indicators of risk in the previous section. Whereas several of the most common location-based model features were recommended by 16 or more participants, the single most common feature for a partner’s likelihood of causing harm was present in only 12 of the participants’ models (gender of the meeting partner). Inconsistency in partner-related model features also manifested in feature weights. With the maximum normalized risk level for any feature being 6, several partner-related features had a maximum risk weight of less than 5. This suggests that qualities of the meeting partner do not impact assessments of risk as much as information about the anticipated meeting location.

Analysis of model features related to a partner’s likelihood of causing harm resulted in three categories of risk, which we unpack in the following subsections: 1) *the ability to personally evaluate a partner before meeting them face-to-face*, 2) *risk associated with demographic traits of the meeting partner*, and 3) *risk identified in data collected privately by AI*.

4.4.1 Detecting Risk Through Ability to Personally Evaluate a Partner Before Meeting Face-to-Face.

Several features suggested by participants for risk detection AI were not directly about their meeting partners, but rather the capacity to collect information about them before face-to-face encounters. Participants explained that reduced access to information before a face-to-face meeting was itself representative of risk because it would raise questions about the validity of conclusions about risk associated with the meeting partner. Thus the more information about a meeting partner, the less risky they would become because participants considered themselves better able predict the behavior and personal qualities of the partner.

The most popular example of this in participants’ models was familiarity with the meeting partner, referring to whether the participant already personally knew the user in question (i.e., met them before discovery on the app). Even though dating apps are intended for discovery of new people, this feature acknowledged the possibility that participants could discover people through the app that they already knew due to geographic proximity - these people would accordingly

represent the lowest risk. For users they did not already know, a similar feature pertained to communication before meeting such as through messaging or phone calls to develop a sense of familiarity. Where these two features differ is when familiarity is developed - prior to discovering the user online or post-profile discovery. Another, less popular example was a feature called "mutual connections", with the understanding that the participant could use a shared acquaintance or friend as a proxy for familiarity with the meeting partner. P2 described personal strategies of using mutual connections and their personal relationships with those connections as a way to discern what a potential meeting partner might be like: "If they're connected with people that have my shared values, then that's a draw. And if they're connected with people that I just know very well, but I'm very different from them in terms of values, they're probably not going to be the best fit for me."

Several women also honed onto profile pages and had strongly-held theories about personal traits that could be inferred through "completeness" of profiles (independent of specific content in the profile). For instance, relatively incomplete profiles - characterized by short personal bio descriptions and missing information in dedicated fields on the profile - were considered an indicator of interest in casual sex, which some participants found risky if they were not personally seeking sexual encounters. P5 described how she makes this inference: "If you can't be bothered to put in a little bit of effort [to properly complete a profile] in order to like, meet your future boyfriend, girlfriend? Like, what's your real end goal for being on this site?"

Table 7. Visualization of model features indicative of ability to personally evaluate a partner before meeting. Participants considered reduced access to information before a face-to-face meeting to increase risk because it jeopardized the accuracy of their assessment of the meeting partner. The features are sorted according to the number of participants that had the feature in their model. The numbers within each cell reflect the maximum weight that the participant gave to the feature, normalized across a 1-6 scale with 1 reflecting the least risk and 6 reflecting the most risk.

Participant	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P12 (g ^{**})	P13	P14	P15	P16	P17	P18	P19	P20
Familiarity with meeting partner(s)	6	5.4		4	4.6	4	3.5							4.3		5	6				6
Profile completeness	6				6	6		6		4.8	5.8										6
Communication before meeting								6		3.5	5.3							4.3			
Able to view profile(s) before meeting												4.9	4.9					4.3	4.8	3.5	
Mutual connections		4.9				6					3.8										
<i>Photo of group activity*</i>					6																
Passive observation of partner									2.7												

*Italicized features exclusively pertain to group events. **P12 made a separate risk detection model for group events.

4.4.2 *Detecting Risk Through Demographic Traits of the Meeting Partner.* All but three participants identified at least one personal quality or demographic trait of meeting partners that they considered reflective of risk. These were qualities that are not expressly indicative of risk, but which participants believed were reliable indicators of one’s propensity to cause harm due to prior personal experiences with people having those traits.

The most common of these was gender of the meeting partner(s), with cisgender men being universally considered higher risk than any other gender identity because of past experiences with “*misogyny*” and physical force used during sex, as well as their general ability to physically overpower women. On the contrary, one transgender participant (P8) considered cisgender *women* as higher risk due to past experiences of emotional harm that came with meeting a group of cisgender women who perceived them as a threat because of their physical stature: “...*It’s people making assumptions based on your appearance. [...] Instead of perceiving physical risk to yourself, you’re thinking like I could be perceived as threatening [to others] or I could be judged negatively.*”

Age difference was also frequently mentioned by participants, although they had different ways of inferring risk through the age of their meeting partner. For example, P1 talked about risk of meeting users under the age of 18 because it could damage their reputation or lead to an arrest: “*If it was a bunch of high schoolers hanging out, I wouldn’t want to do that. If I was a man, and I accidentally showed up to one of these events, and there’s like, a bunch of high schoolers, I would feel at risk of being arrested or something. Or being publicly shamed for accidentally rolling up on a bunch of high schoolers.*” P2, on the other hand, connected this feature with gender and expressed being more concerned with meeting men who are approximately middle-aged because they may have more physical strength to sexually assault them than older adults. As they put it: “*I’m not going to be intimidated to meet up with a man that’s 80 years old, versus somebody that’s maybe 30 years old and single.*”

Table 8. Visualization of model features evaluating demographic traits of meeting partner. The demographic traits considered indicative of risk were deeply personal and varied considerably across participants. The features are sorted according to the number of participants that had the feature in their model. The numbers within each cell reflect the maximum weight that the participant gave to the feature, normalized across a 1-6 scale with 1 reflecting the least risk and 6 reflecting the most risk.

Participant	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P12 (g*)	P13	P14	P15	P16	P17	P18	P19	P20
Gender	4.9	4.6		6		6		4.5		4.8				4.9	6	4	6	4.9			4.8
Age difference		4.3	4.6	6				6	6					3.8							3.5
Spiritual values		4.3	6																		
Education level		5.4	6																		
Shared interests							6														
Political affiliation										5.4					5.4						
Shared student status														4.3			3				
Marital status		4.9																			
Cultural background		4.9																			
Mental health																	4.8				
Personality					4.6																

*P12 made a separate model specifically pertaining to group based events.

Beyond gender and age difference, there was very little commonality across models for other demographic traits that participants associated with risk. This is likely because such traits were informed by deeply personal experiences with risk and safety that may not have been shared between participants. For instance, P3 felt safer around other Muslims due to shared values and past experiences of abuse based on her religious values from non-Muslim meeting partners and thus considered anyone who is not Muslim to pose maximum risk to her safety. The notion of inferring "values" came up with other demographic traits as well, such as political affiliation in models from P10 and P15. They considered some political views to be strong indicators of how someone treats other people, such as through their views on abortion and immigration, and believed one's political party was a suitable proxy for inferring their stance on a variety of human rights-related issues that might cause them to disregard the safety of other dating app users. In their words: "If there's something that really I would disagree with, politically or socio-politically, so they are sort of like human values, just being a respectful human being, a kind human." P2's feature for cultural background had a similar intent, which vaguely sought to evaluate whether a user's "culture" (described through a mix of religion, political views, and geographic area of upbringing) had regressive views about women.

Most of the model features in this category would be enabled through users consciously providing personal data to risk detection AI through private profiles (see section 4.2 for an explanation of this and other means for using to contribute training data). Yet for some reasons participants were vague or uncertain about how risk detection AI would feasibly compute them. This applies to cultural background, as already mentioned, as well as mental health issues and personality because the constructs themselves remained ill-defined. In these cases participants deferred to descriptions of how they personally infer mental health and personality through messaging conversations and profile reviews, with the assumption that risk detection AI could similarly evaluate such traits through scanning interactions between users.

4.4.3 *Detecting Risk Through Data Collected Privately by AI.* Other model features stemmed from possibilities for new data that risk detection AI may have available that normal users would not be able to access directly. These features would be informed by data sources from outside the dating app, as well as data privately provided to risk detection AI by users as described in section 4.2. Some participants described this data as particularly valuable because it would be immune to manipulation from the respective user, unlike profile pages and messaging interactions which could include lies and exaggerated content.

Table 9. Visualization of model features informed by data collected privately by AI. The features are sorted according to the number of participants that had the feature in their model. The numbers within each cell reflect the maximum weight that the participant gave to the feature, normalized across a 1-6 scale with 1 reflecting the least risk and 6 reflecting the most risk.

Participant	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P12 (g**)	P13	P14	P15	P16	P17	P18	P19	P20
Criminal record		5.7		6	6						6	5.4		6	3.5	5	6			4.3	3.5
Reputation of meeting partner(s)	6						5	5.4			5.8			4.9							
Face-to-face meeting history								3.8			2.5										
Criminal record in group*											5										

*Italicized features exclusively pertain to group events. **P12 made a separate risk detection model for group events.

The most common example in participants' models was criminal record of the meeting partner, which participants expected risk detection AI to access through government sources. This would necessitate users providing their real identities to the risk detection AI through private profiles. The prevalence of this feature in participants' models is not an indication that they believe most users do have criminal records, but rather an opportunity to identify presumably rare convictions for violent crimes or sexual offenses. Some participants gave relatively small risk weights for this feature (e.g., a maximum of 3.5 out of 6) based on the assumption that the most common crimes are not applicable to risk of interpersonal harm and thus should not affect overall risk scores too strongly.

Other features in this category, particularly reputation of meeting partners and face-to-face meeting history, would be the direct result of new data collected from users through reviews of their face-to-face meeting partners (see section 4.2 for discussion of user reviews). Face-to-face meeting history would be based on a count of the number of users one has met face-to-face from the app, which would be confirmed by their face-to-face meeting partners. A lower number of prior face-to-face meetings would increase risk because of the uncertainties it would present about the user's intentions for app-use or if their profile was even real. Reputation scores would be based on reviews of the partner's behavior during the face-to-face meeting (envisioned as a percentage by some participants). This feature is different from the traditional reporting/blocking features in dating apps today because reputation scores would not be based simply on whether the user actually caused harm (which may result in being banned from the platform), but whether they have made past partners feel uncomfortable during face-to-face meetings.

5 Limitations

There are some limitations to the method deserving of note. Despite participants exhibiting rapid comprehension of the model building exercise, there is uncertainty around how their ideas would translate to more practical machine learning models. Weak points of the method in retrospect are unknown feasibility (and legality) of some model features given that participants could propose features without any regard to availability of data/inputs. This limitation was further exemplified in some participants' struggles to operationalize some features beyond vague references (e.g., personality traits). While legitimate model features for personality may exist, they may not match participants' conceptualizations of this ill-defined construct. On a similar note, participants' risk detection models may have manifested differently if they were more acutely aware of existing risk detection AI implementations and their criticisms (e.g., [100]).

Furthermore, we did not collect demographic details around duration of dating app-use by our participants. While the substance of our participants' experiences pertaining to risk and harm were unpacked during interviews, quantitative metrics such as duration and frequency of dating app-use could have lent additional context to our findings pertinent to future work seeking to operationalize new forms of risk detection AI for the broader userbase. In addition, because the notion of user-personalized risk detection models became apparent in the data analysis phase we were not able to ask participants directly about the distinction between a universal risk detection model for all dating app users and user-customizable models.

Lastly we consider the implications of our interface-based elicitation of risk detection models. We motivated model brainstorming with profile mock ups exhibiting a risk indicator (low, medium, or high), which participants easily understood and engaged with. However this limited opportunity for stakeholders to design how risk detection AI decisions should be conveyed to users. They tended to assume the low/medium/high risk indicator was an unchangeable design pattern despite occasional vocalized struggles to connect numerical risk scores to these risk categories. In retrospect we are skeptical whether the risk indicators in the profile mock ups were essential to model building; the profiles themselves sans-risk indicators may have been sufficient. In the future we would refrain from “pre-designing” AI output in the respective interface unless confirmed necessary for comprehension.

6 Discussion

In response to alarming rates of dating app-facilitated sexual violence against women and other marginalized groups [3, 4] and calls for increased stakeholder participation in design of risk detection AI, our work involves woman-identifying dating app users in envisioning interactions with risk detection AI and designing risk detection models pursuant to those interactions.

Findings show that women anticipate using risk detection AI as an assistant, rather than replacement, to their manually conducted safety strategies. This reinforces arguments in the literature highlighting the need for algorithmic systems that augment, rather than replace, human decision making [109]. Anticipated interactions with risk detection AI following from this conceptualization include donating personal data to the AI to help it understand how its user personally assesses risk as well as reviewing explanations of the AI’s risk computation to check for inconsistencies with one’s personal strategies. In this section we reflect on the implications of our study for risk detection AI research and design in three areas: distinguishing different types of risk detection AI and their relative research progress, future avenues for participatory AI design to actualize personalizable risk detection, and ethical concerns with personalizable risk detection AI.

6.1 Distinguishing Risk from Harm: Implications of Location-Based Risk Detection AI

The concepts of “risk” and “harm” tend to be conflated in risk detection AI literature [100]. For example, terminology around risk detection AI has been applied to models for identifying actualized harm such as unsolicited nude imagery [121], harassing messages [10], and cyberbullying [72]. It also includes detection of harm attempts such as child grooming and sex trafficking solicitations [14, 100, 129]. A common theme amongst these prior examples is a focus on the perpetrator: identifying harmful acts committed by the perpetrator or being planned by the perpetrator.

On the contrary, our study found that women conceptualize risk beyond the likelihood of a specific person causing harm. Risk also encapsulates their personal capacity to maintain safety. In our study’s context of meeting dating app users face-to-face this manifested through the ability to seek help from others at date locations, physically eject from unsafe situations, and foresee danger associated with particular face-to-face meeting locations. Our findings show a clear distinction between person-based and location-based indicators of risk, with location-based features arguably being more important given how consistently participants advocated for them, and how heavily weighted they were in their risk detection models. For example, the three most popular features across all participants’ risk detection models were the presence of bystanders at a face-to-face meeting location, familiarity with the face-to-face meeting location, and the number of people being met at the location.

We thus urge future research to better distinguish between these three forms of detection AI, specifically: detection of actualized harm (harm detection AI), detection of attempted or potential harm associated with a specific person (interpersonal risk detection AI), and detection of one’s

personal capacity to manage risk and harm in the context within which it could occur (location-based risk detection AI). Distinguishing these types of AI is more than a matter of semantics; it can also elucidate relative gaps in research and design. For instance, while literature has documented technical and human-centered design progress with harm detection AI and interpersonal risk detection AI (e.g., [71, 100]), there is an absence of attention to AI for detecting location-based risk.

The importance of location-based risk detection AI, and human-centered research around it, becomes evident when situated amongst other technologies proposed or studied in HCI literature for addressing interpersonal harm in physical environments. These include mobile apps that afford women with services such as GPS tracking [136] and safety monitoring [103] as well as crowdsourced safe routes and safety alerts to help avoid unsafe areas [5, 108]. Location-based risk detection AI can add scale and automation to these other risk mitigation features, and also address limitations of related tools such as panic buttons [68] that have been critiqued for being reactive to - rather than preventative of - harm.

6.2 Implications of Subjective Risk Detection on Participatory AI

In addition to broadening the scope of risk detection AI, our findings also emphasize that risk is a subjective concept, reflective of personal experience and one's *perceived* capacity to foresee and manage unsafe situations. Our participants varied significantly in which traits of a meeting partner they considered indicative of risk, such as the religion of a meeting partner, political affiliation, education background, and age difference. Accordingly, participants expected risk detection AI to learn and adapt to their personal safety strategies. This would imply that a singular, one-size-fits-all risk detection model in dating apps (or any other social platform) would inevitably fail to meet the needs of all stakeholder groups because it would be forced to conceptualize risk in a way that aligns with only a subset of any given population.

Drawing from the principles of feminist HCI [17], the findings of our study could be used to advocate for pluralist design of risk detection. We connect our study with prior work in three areas to chart directions for future research into pluralist risk detection AI: 1) improving ground truth for risk, 2) data donation for model training, and 3) participatory AI model building.

A recurring criticism of risk detection AI is a reliance on external annotators who label data indicative of risk or harm regardless of whether they have personally experienced the respective behavior [9, 71, 72]. This can result in inconsistent ground truth for harm that does not match the understanding of victims and thus expose AI to misidentification of harm. Like prior work [71], we advocate for involvement of stakeholders who at risk of (or already been victims of) the respective harm in data labeling, albeit our reasoning is not in pursuit of a more consistent ground truth - we actually argue the opposite. Given our findings that identification of risk is subjective and personal, having beneficiaries/users of risk detection AI provide ground truth for risk can enable users to personalize their risk detection. In other words, instead of converging on a singular ground truth for risk, empowering users as risk annotators can be a way to enable multiple ground truths that map to each user's personal understanding of risk.

Data donation has emerged as a promising pathway through which end-users can get involved in annotating risk-relevant data, and also improving the quality of datasets that risk detection AI models are trained on [16, 99]. Importantly, our study's findings also indicate that user motivations for donating personal data go beyond altruistic motives of societal impact as found in prior work [44, 112]. Our participants talked extensively about desire and willingness to contribute personal data to risk detection AI with the anticipation that it would directly benefit them through preparing the AI to assess risk according to their subjective safety strategies. Examples from our study included a secondary profile page viewable only by risk detection AI for providing personal information as training data. Prior data donation research has produced functional applications [57, 99, 100] and

empirical insight [16, 56, 58, 65, 80, 137] for supporting users in uploading sensitive and personal data. This prior work can serve as a foundation for design of these imagined “secondary profiles” for users to prepare their risk detection AI.

Our participants imagined the donation of personal data as a way to enable every user’s personal risk detection AI - in effect providing a corpus of potential data points to be added, removed, and modified in each user’s customized risk detection model. While we used prior research on participatory model building [76–79] as inspiration in our study’s method, the findings suggest that such methods could also be incorporated directly into social platform interfaces to support intuitive modification of risk detection models by users with limited AI literacy. For instance, methods of visualizing model training such as paired comparisons or weighting of directly explainable features [78] could be assessed in future usability research of participatory risk detection AI model personalization interfaces.

6.3 Problematising Subjective Risk Detection AI

While the aforementioned approaches can pave the way towards personalized risk detection AI, the potentially adverse consequences of such AI should be considered. The manually-performed risk assessment strategies described by our participants often leveraged indicators of risk with dubious scientific backing, and in some cases could be construed as unfairly biased towards particular demographics. This is most poignant through examples from our participants such as assuming people of certain religions are more dangerous than others, or that people with a high school education pose more risk than college educated users.

Allowing unfettered personalization of risk detection AI could encourage, affirm, and amplify harmful stereotypes against marginalized groups. HCI literature has described this reinforcing of harmful stereotypes as algorithmic symbolic annihilation [12]. Similarly, Karizat et al.[67] report algorithmic representational harms on social platforms through which users’ identities are suppressed and misrepresented by algorithms. In the online dating context in particular, research has extensively explored identity-based harms that users of marginalized groups try to manage. This includes strategic disclosure by transgender users [51] and users with disabilities [97], as well careful presentation on apps for men-seeking-men to prevent deduction of personal identities that could incur physical harm [19]. In this light, personalizable risk detection AI could actually increase risk to users through allowing and affirming villainization of marginalized identity traits. This could, in turn, necessitate even more extensive manual effort to mitigate risk associated with users, meeting locations, and now algorithms. The algorithmic harm literature gives numerous examples of user strategies to combat perceived adverse algorithmic processes, including identity modulation and flattening [43, 47, 128] and outright leaving the applicable platform [42] - outcomes that certainly run contrary to the intent of risk detection AI.

Future research and design for risk detection AI needs to consider guardrails or limitations on user-personalized models to mitigate adverse impact. This may include 1) providing a limited range of model features that users can personalize, 2) using AI to generate counterarguments towards user-selected features that are potentially biased, or 3) avoiding user personalization and instead incorporating ample explanation of a universal risk detection model’s decision making so that a user can decide if or how to incorporate the model’s conclusion into their own risk assessment.

7 Conclusion

In response to the prevalence of interpersonal harms against women across online and physical modalities, risk detection AI implementations have grown in popularity across social computing platforms as a scalable approach to mitigating harm. While there is extensive literature into the computational performance of “after-the-fact” risk detection models, there is a relative gap in

knowledge regarding how user interactions with risk detection AI should be designed to successfully keep users safe - what prior work has described as the difference between risk detection and risk mitigation [120]. To address this gap, this paper presented an interview study containing participatory risk detection model building activities with women (n=20) about how they envision interacting with risk detection AI in dating apps to mitigate risk of harm associated with meeting other users face-to-face, and how risk detection models can be designed to realize those interactions.

Findings highlight expectations of interacting with risk detection AI as a partner that assists women in their manually-practiced risk assessment strategies. Anticipated interactions with risk detection AI towards this goal include training the AI to understand their personal and subjective risk assessment strategies, inspecting how the AI arrives at its risk conclusions to determine if it is “thinking” about risk in the same way they do, and informing their own ongoing risk assessment with new data collected by the AI. The findings suggest that future work should explore the notion of personalizable risk detection models that define and assess risk according to the personal preferences of each user, while also considering ethical implications such as the reinforcing of problematic stereotypes about who poses risk of harm.

Acknowledgments

This work is partially supported by the U.S. National Science Foundation under Grant No. 2211896 and Grant No. 2339431. We thank Hanan Aljasim and Caroline Bull for their efforts regarding participant recruitment and data collection for this study.

References

- [1] 2021. Tinder introduces are you sure?, an industry-first feature that is stopping harassment before it starts. <https://www.tinderpressroom.com/2021-05-20-Tinder-Introduces-Are-You-Sure-,-an-Industry-First-Feature-That-is-Stopping-Harassment-Before-It-Starts>
- [2] 2022. *Centers for Disease Control and Prevention* (Jun 2022). <https://www.cdc.gov/violenceprevention/sexualviolence/fastfact.html>
- [3] 2022. More under 20s sexually assaulted after meeting offenders on dating sites. *National Crime Agency* (Feb 2022). <https://www.nationalcrimeagency.gov.uk/news/more-under-20s-sexually-assaulted-after-meeting-offenders-on-dating-sites>
- [4] National Crime Agency. 2016. Emerging new threat in online dating: Initial trends in internet dating-initiated serious sexual assaults. (2016).
- [5] Syed Ishtiaque Ahmed, Steven J. Jackson, Nova Ahmed, Hasan Shahid Ferdous, Md. Rashidujjaman Rifat, A.S.M Rizvi, Shamir Ahmed, and Rifat Sabbir Mansur. 2014. Protibadi: A platform for fighting sexual harassment in urban Bangladesh. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Toronto, Ontario, Canada) (*CHI '14*). Association for Computing Machinery, New York, NY, USA, 2695–2704. <https://doi.org/10.1145/2556288.2557376>
- [6] Kath Albury, Christopher Dietzel, Tinonee Pym, Son Vivienne, and Teddy Cook. 2021. Not your unicorn: trans dating app users’ negotiations of personal safety and sexual health. *Health Sociology Review* 30, 1 (2021), 72–86. <https://doi.org/10.1080/14461242.2020.1851610> arXiv:<https://doi.org/10.1080/14461242.2020.1851610> PMID: 33622202.
- [7] Abdullah X. Ali, Meredith Ringel Morris, and Jacob O. Wobbrock. 2018. Crowdsourcing Similarity Judgments for Agreement Analysis in End-User Elicitation Studies. In *Proceedings of the 31st Annual ACM Symposium on User Interface Software and Technology* (Berlin, Germany) (*UIST '18*). Association for Computing Machinery, New York, NY, USA, 177–188. <https://doi.org/10.1145/3242587.3242621>
- [8] Mohammed Eunus Ali, Shabnam Basera Rishta, Lazima Ansari, Tanzima Hashem, and Ahamad Imtiaz Khan. 2015. SafeStreet: Empowering Women against Street Harassment Using a Privacy-Aware Location Based Application. , *Article 24* (2015), 4 pages. <https://doi.org/10.1145/2737856.2737870>
- [9] Shiza Ali, Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Joshua Gracie, Munmun De Choudhury, Pamela J. Wisniewski, and Gianluca Stringhini. 2022. Understanding the Digital Lives of Youth: Analyzing Media Shared within Safe Versus Unsafe Private Conversations on Instagram. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (*CHI '22*). Association for Computing Machinery, New York, NY, USA, Article 148, 14 pages. <https://doi.org/10.1145/3491102.3501969>

- [10] Ashwaq Alsoubai, Xavier V. Caddle, Ryan Doherty, Alexandra Taylor Koehler, Estefania Sanchez, Munmun De Choudhury, and Pamela J. Wisniewski. 2022. MOSafely, Is That Sus? A Youth-Centric Online Risk Assessment Dashboard. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing (Virtual Event, Taiwan) (CSCW'22 Companion)*. Association for Computing Machinery, New York, NY, USA, 197–200. <https://doi.org/10.1145/3500868.3559710>
- [11] Oscar Alvarado and Annika Waern. 2018. Towards Algorithmic Experience: Initial Efforts for Social Media Contexts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (Montreal QC, Canada) (CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–12. <https://doi.org/10.1145/3173574.3173860>
- [12] Nazanin Andalibi and Patricia Garcia. 2021. Sensemaking and coping after pregnancy loss: the seeking and disruption of emotional validation online. *Proceedings of the ACM on Human-Computer Interaction* 5, CSCW1 (2021), 1–32.
- [13] Nazanin Andalibi, Oliver L. Haimson, Munmun De Choudhury, and Andrea Forte. 2016. Understanding Social Media Disclosures of Sexual Abuse Through the Lenses of Support Seeking and Anonymity. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems (San Jose, California, USA) (CHI '16)*. Association for Computing Machinery, New York, NY, USA, 3906–3918. <https://doi.org/10.1145/2858036.2858096>
- [14] Philip Anderson, Zheming Zuo, Longzhi Yang, and Yanpeng Qu. 2019. An Intelligent Online Grooming Detection System Using AI Technologies. In *2019 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE)*. 1–6. <https://doi.org/10.1109/FUZZ-IEEE.2019.8858973>
- [15] Anuoluwapo Abosede Durokifa Andrew Enaifoghe, Melita Dlelana and Nomaswazi P. Dlamini. 2021. The Prevalence of Gender-Based Violence against Women in South Africa : A Call for Action. *African Journal of Gender, Society and Development (formerly Journal of Gender, Information and Development in Africa)* 10, 1 (2021), 117–146. <https://doi.org/10.31920/2634-3622/2021/v10n1a6> arXiv:<https://journals.co.za/doi/pdf/10.31920/2634-3622/2021/v10n1a6>
- [16] Karla Badillo-Urquiola, Zachary Shea, Zainab Agha, Irina Lediaeva, and Pamela Wisniewski. 2021. Conducting Risky Research with Teens: Co-Designing for the Ethical Treatment and Protection of Adolescents. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW3, Article 231 (jan 2021), 46 pages. <https://doi.org/10.1145/3432930>
- [17] Shaowen Bardzell. 2010. Feminist HCI: Taking Stock and Outlining an Agenda for Design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Atlanta, Georgia, USA) (CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1301–1310. <https://doi.org/10.1145/1753326.1753521>
- [18] Eric PS Baumer. 2017. Toward human-centered algorithm design. *Big Data & Society* 4, 2 (2017), 2053951717718854. <https://doi.org/10.1177/2053951717718854> arXiv:<https://doi.org/10.1177/2053951717718854>
- [19] Jeremy Birnholtz, Shruta Rawat, Richa Vashista, Dicky Baruah, Alpna Dange, and Anne-Marie Boyer. 2020. Layers of Marginality: An Exploration of Visibility, Impressions, and Cultural Context on Geospatial Apps for Men Who Have Sex With Men in Mumbai, India. *Social Media + Society* 6, 2 (2020), 2056305120913995. <https://doi.org/10.1177/2056305120913995> arXiv:<https://doi.org/10.1177/2056305120913995>
- [20] Jeremy Birnholtz, Irina Shklovski, Mark Handel, and Eran Toch. 2015. Let's Talk About Sex (Apps), CSCW. In *Proceedings of the 18th ACM Conference Companion on Computer Supported Cooperative Work & Social Computing (Vancouver, BC, Canada) (CSCW'15 Companion)*. Association for Computing Machinery, New York, NY, USA, 283–288. <https://doi.org/10.1145/2685553.2685557>
- [21] Courtney Blackwell, Jeremy Birnholtz, and Charles Abbott. 2015. Seeing and being seen: Co-situation and impression formation using Grindr, a location-aware gay dating app. *New Media & Society* 17, 7 (2015), 1117–1136. <https://doi.org/10.1177/1461444814521595> arXiv:<https://doi.org/10.1177/1461444814521595>
- [22] Jan Blom, Divya Viswanathan, Mirjana Spasojevic, Janet Go, Karthik Acharya, and Robert Athonius. 2010. Fear and the City: Role of Mobile Services in Harnessing Safety and Security in Urban Use Contexts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (Atlanta, Georgia, USA) (CHI '10)*. Association for Computing Machinery, New York, NY, USA, 1841–1850. <https://doi.org/10.1145/1753326.1753602>
- [23] Elizabeth Bondi, Lily Xu, Diana Acosta-Navas, and Jackson A. Killian. 2021. Envisioning Communities: A Participatory Approach Towards AI for Social Good. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society (Virtual Event, USA) (AI/ES '21)*. Association for Computing Machinery, New York, NY, USA, 425–436. <https://doi.org/10.1145/3461702.3462612>
- [24] Robert Bowman, Camille Nadal, Kellie Morrissey, Anja Thieme, and Gavin Doherty. 2023. Using Thematic Analysis in Healthcare HCI at CHI: A Scoping Review. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (-conf-loc-, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 491, 18 pages. <https://doi.org/10.1145/3544548.3581203>
- [25] Tone Bratteteig and Ina Wagner. 2016. Unpacking the Notion of Participation in Participatory Design. *Comput. Supported Coop. Work* 25, 6 (dec 2016), 425–475. <https://doi.org/10.1007/s10606-016-9259-4>
- [26] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology* 3, 2 (2006), 77–101. <https://doi.org/10.1191/1478088706qp0630a> arXiv:<https://www.tandfonline.com/doi/pdf/10.1191/1478088706qp0630a>

- [27] Virginia Braun and Victoria Clarke. 2021. Thematic Analysis: A Practical Guide. (2021), 1–376.
- [28] Virginia Braun, Victoria Clarke, Nikki Hayfield, and Gareth Terry. 2019. *Thematic Analysis*. Springer Singapore, Singapore, 843–860. https://doi.org/10.1007/978-981-10-5251-4_103
- [29] Denton Callander, Martin Holt, and Christy E. Newman. 2016. ‘Not everyone’s gonna like me’: Accounting for race and racism in sex and dating web services for gay and bisexual men. *Ethnicities* 16, 1 (2016), 3–21. <https://doi.org/10.1177/1468796815581428> arXiv:<https://doi.org/10.1177/1468796815581428>
- [30] Karen A Campbell, Elizabeth Orr, Pamela Durepos, Linda Nguyen, Lin Li, Carly Whitmore, Paige Gehrke, Leslie Graham, and Susan M Jack. 2021. Reflexive thematic analysis for applied qualitative health research. *The Qualitative Report* 26, 6 (2021), 2011–2028.
- [31] Vanessa Centelles, R achael A. Powers, and Jr Richard K. Moule. 2021. An Examination of Location-Based Real-Time Dating Application Infrastructure, Profile Features, and Cybervictimization. *Social Media + Society* 7, 3 (2021), 20563051211043218. <https://doi.org/10.1177/20563051211043218> arXiv:<https://doi.org/10.1177/20563051211043218>
- [32] Janet X. Chen, Allison McDonald, Yixin Zou, Emily Tseng, Kevin A Roundy, Acar Tamersoy, Florian Schaub, Thomas Ristenpart, and Nicola Dell. 2022. Trauma-Informed Computing: Towards Safer Technology Experiences for All. In *CHI Conference on Human Factors in Computing Systems*. ACM, New Orleans LA USA, 1–20. <https://doi.org/10.1145/3491102.3517475>
- [33] Lu Cheng, Jundong Li, Yasin N. Silva, Deborah L. Hall, and Huan Liu. 2019. XBully: Cyberbullying Detection within a Multi-Modal Context. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining (Melbourne VIC, Australia) (WSDM ’19)*. Association for Computing Machinery, New York, NY, USA, 339–347. <https://doi.org/10.1145/3289600.3291037>
- [34] Edmond Pui Hang Choi, Janet Yuen Ha Wong, and Daniel Yee Tak Fong. 2018. An Emerging Risk Factor of Sexual Abuse: The Use of Smartphone Dating Applications. *Sexual Abuse* 30, 4 (2018), 343–366. <https://doi.org/10.1177/1079063216672168> arXiv:<https://doi.org/10.1177/1079063216672168>
- [35] Sasha Costanza-Chock. 2018. Design Justice, A.I., and Escape from the Matrix of Domination. *Journal of Design and Science* (jul 16 2018). <https://jods.mitpress.mit.edu/pub/costanza-chock>.
- [36] Jeffrey Dastin. 2018. Amazon scraps secret AI recruiting tool that showed bias against women. In *Ethics of data and analytics*. Auerbach Publications, 296–299.
- [37] Isha Datey, Hanan Khalid Aljasim, and Douglas Zytka. 2022. Repurposing AI in Dating Apps to Augment Women’s Strategies for Assessing Risk of Harm. In *Companion Publication of the 2022 Conference on Computer Supported Cooperative Work and Social Computing (Virtual Event, Taiwan) (CSCW’22 Companion)*. Association for Computing Machinery, New York, NY, USA, 150–154. <https://doi.org/10.1145/3500868.3559472>
- [38] Walter S DeKeseredy. 2020. Understanding the harms of pornography: The contributions of social scientific knowledge. *Culture Reframed* (2020), 1–16.
- [39] Bharat H Desai and Moumita Mandal. 2021. Role of climate change in exacerbating sexual and gender-based violence against women: A new challenge for international law. *Environmental Policy and Law* 51, 3 (2021), 137–157.
- [40] Thomas Deselaers, Lexi Pimenidis, and Hermann Ney. 2008. Bag-of-visual-words models for adult image classification and filtering. In *2008 19th International Conference on Pattern Recognition*. 1–4. <https://doi.org/10.1109/ICPR.2008.4761366>
- [41] Prema Dev, Jessica Medina, Zainab Agha, Munmun De Choudhury, Afsaneh Razi, and Pamela J. Wisniewski. 2022. From Ignoring Strangers’ Solicitations to Mutual Sexting with Friends: Understanding Youth’s Online Sexual Risks in Instagram Private Conversations (CSCW’22 Companion). Association for Computing Machinery, New York, NY, USA, 94–97. <https://doi.org/10.1145/3500868.3559469>
- [42] Michael Ann DeVito. 2022. How transfeminine TikTok creators navigate the algorithmic trap of visibility via folk theorization. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–31.
- [43] Michael A DeVito, Ashley Marie Walker, and Jeremy Birnholtz. 2018. ‘Too Gay for Facebook’ Presenting LGBTQ+ Identity Throughout the Personal Social Media Ecosystem. *Proceedings of the ACM on Human-Computer Interaction* 2, CSCW (2018), 1–23.
- [44] Daniel Diethel, Jasmin Niess, Carolin Stellmacher, Evropi Stefanidi, and Johannes Sch onig. 2021. Sharing Heartbeats: Motivations of Citizen Scientists in Times of Crises. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Yokohama</city>, <country>Japan</country>, </conf-loc>) (CHI ’21). Association for Computing Machinery, New York, NY, USA, Article 650, 15 pages. <https://doi.org/10.1145/3411764.3445665>
- [45] Jill P. Dimond, Michaelanne Dye, Daphne Larose, and Amy S. Bruckman. 2013. Hollaback! The Role of Storytelling Online in a Social Movement Organization. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (San Antonio, Texas, USA) (CSCW ’13)*. Association for Computing Machinery, New York, NY, USA, 477–490. <https://doi.org/10.1145/2441776.2441831>

- [46] Caitlin H Douglass, Cassandra JC Wright, Angela C Davis, and Megan SC Lim. 2018. Correlates of in-person and technology-facilitated sexual harassment from an online survey among young Australians. *Sexual health* 15, 4 (2018), 361–365. <https://doi.org/10.1071/SH17208>. PMID: 29852924.
- [47] Stefanie Duguay. 2017. *Identity modulation in networked publics: Queer women's participation and representation on Tinder, Instagram, and Vine*. Ph. D. Dissertation. Queensland University of Technology.
- [48] Stefanie Duguay, Jean Burgess, and Nicolas Suzor. 2020. Queer women's experiences of patchwork platform governance on Tinder, Instagram, and Vine. *Convergence* 26, 2 (2020), 237–252. <https://doi.org/10.1177/1354856518781530> arXiv:<https://doi.org/10.1177/1354856518781530>
- [49] Stine Eckert and Jade Metzger-Riftkin. 2020. *Doxxing*. John Wiley & Sons, Ltd, 1–5. <https://doi.org/10.1002/9781119429128.iegmc009> arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781119429128.iegmc009>
- [50] Nicole Ellison, Rebecca Heino, and Jennifer Gibbs. 2006. Managing Impressions Online: Self-Presentation Processes in the Online Dating Environment. *Journal of Computer-Mediated Communication* 11, 2 (2006), 415–441. <https://doi.org/10.1111/j.1083-6101.2006.00020.x> arXiv:<https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1083-6101.2006.00020.x>
- [51] Julia R. Fernandez and Jeremy Birnholtz. 2019. "I Don't Want Them to Not Know": Investigating Decisions to Disclose Transgender Identity on Dating Platforms. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 226 (nov 2019), 21 pages. <https://doi.org/10.1145/3359328>
- [52] Guo Freeman, Samaneh Zamanifard, Divine Maloney, and Dane Acena. 2022. Disturbing the Peace: Experiencing and Mitigating Emerging Harassment in Social Virtual Reality. *Proceedings of the ACM on Human-Computer Interaction* 6 (3 2022), 1–30. Issue CSCW1. <https://doi.org/10.1145/3512932>
- [53] Arijit Ghosh Chowdhury, Ramit Sawhney, Puneet Mathur, Debanjan Mahata, and Rajiv Ratn Shah. 2019. Speak up, Fight Back! Detection of Social Media Disclosures of Sexual Harassment. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Student Research Workshop*. Association for Computational Linguistics, Minneapolis, Minnesota, 136–146. <https://doi.org/10.18653/v1/N19-3018>
- [54] Jennifer L. Gibbs, Nicole B. Ellison, and Chih-Hui Lai. 2011. First Comes Love, Then Comes Google: An Investigation of Uncertainty Reduction Strategies and Self-Disclosure in Online Dating. *Communication Research* 38, 1 (2011), 70–100. <https://doi.org/10.1177/0093650210377091> arXiv:<https://doi.org/10.1177/0093650210377091>
- [55] Louisa Gilbert, Aaron L. Sarvet, Melanie Wall, Kate Walsh, Leigh Reardon, Patrick Wilson, John Santelli, Shamus Khan, Martie Thompson, Jennifer S. Hirsch, and Claude A. Mellins. 2019. Situational Contexts and Risk Factors Associated with Incapacitated and Nonincapacitated Sexual Assaults Among College Women. *Journal of Women's Health* 28, 2 (2019), 185–193. <https://doi.org/10.1089/jwh.2018.7191> arXiv:<https://doi.org/10.1089/jwh.2018.7191> PMID: 30481099.
- [56] Alejandra Gomez Ortega, Jacky Bourgeois, Wiebke Toussaint Hutiri, and Gerd Kortuem. 2023. Beyond data transactions: a framework for meaningfully informed data donation. *AI & SOCIETY* (2023), 1–18. <https://doi.org/10.1007/s00146-023-01755-5>
- [57] Alejandra Gomez Ortega, Jacky Bourgeois, and Gerd Kortuem. 2022. Reconstructing Intimate Contexts through Data Donation: A Case Study in Menstrual Tracking Technologies. In *Nordic Human-Computer Interaction Conference (Aarhus, Denmark) (NordicCHI '22)*. Association for Computing Machinery, New York, NY, USA, Article 8, 12 pages. <https://doi.org/10.1145/3546155.3546646>
- [58] Alejandra Gómez Ortega, Jacky Bourgeois, and Gerd Kortuem. 2023. What is Sensitive About (Sensitive) Data? Characterizing Sensitivity and Intimacy with Google Assistant Users. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (<conf-loc>, <city>Hamburg</city>, <country>Germany</country>, </conf-loc>) (CHI '23). Association for Computing Machinery, New York, NY, USA, Article 586, 16 pages. <https://doi.org/10.1145/3544548.3581164>
- [59] Naeemul Hassan, Amrit Poudel, Jason Hale, Claire Hubacek, Khandakar Tasnim Huq, Shubhra Kanti Karmaker Santu, and Syed Ishtiaque Ahmed. 2019. Towards Automated Sexual Violence Report Tracking. arXiv. <https://doi.org/10.48550/ARXIV.1911.06961>
- [60] Nicola Henry and Anastasia Powell. 2015. Embodied Harms: Gender, Shame, and Technology-Facilitated Sexual Violence. *Violence Against Women* 21, 6 (2015), 758–779. <https://doi.org/10.1177/1077801215576581> arXiv:<https://doi.org/10.1177/1077801215576581> PMID: 25827609.
- [61] Nicola Henry, Anastasia Powell, and Asher Flynn. 2018. AI can now create fake porn, making revenge porn even more complicated. *The Conversation* 28 (2018).
- [62] Joey Chiao-Yin Hsiao and Tawanna R. Dillahunt. 2017. People-Nearby Applications: How Newcomers Move Their Relationships Offline and Develop Social and Cultural Capital. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) (CSCW '17). Association for Computing Machinery, New York, NY, USA, 26–40. <https://doi.org/10.1145/2998181.2998280>
- [63] Larke N Huang, Rebecca Flatow, Tenly Biggs, Sara Afayee, Kelley Smith, Thomas Clark, and Mary Blake. 2014. SAMHSA's Concept of Trauma and Guidance for a Trauma-Informed Approach. (2014).

- [64] Yi Huang and Adams Wai Kin Kong. 2016. Using a CNN ensemble for detecting pornographic and upskirt images. In *2016 IEEE 8th International Conference on Biometrics Theory, Applications and Systems (BTAS)*. 1–7. <https://doi.org/10.1109/BTAS.2016.7791207>
- [65] Jack Jamieson and Naomi Yamashita. 2023. Escaping the Walled Garden? User Perspectives of Control in Data Portability for Social Media. *Proc. ACM Hum.-Comput. Interact.* 7, CSCW2, Article 339 (oct 2023), 27 pages. <https://doi.org/10.1145/3610188>
- [66] Younes Karimi, Anna Squicciarini, and Shomir Wilson. 2022. Automated Detection of Doxing on Twitter. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 276 (nov 2022), 24 pages. <https://doi.org/10.1145/3555167>
- [67] Nadia Karizat, Dan Delmonaco, Motahhare Eslami, and Nazanin Andalibi. 2021. Algorithmic folk theories and identity: How TikTok users co-produce Knowledge of identity and engage in algorithmic resistance. *Proceedings of the ACM on human-computer interaction* 5, CSCW2 (2021), 1–44.
- [68] Naveena Karusala and Neha Kumar. 2017. Women’s Safety in Public Spaces: Examining the Efficacy of Panic Buttons in New Delhi. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) (*CHI ’17*). Association for Computing Machinery, New York, NY, USA, 3340–3351. <https://doi.org/10.1145/3025453.3025532>
- [69] Shamus Khan, Jennifer Hirsch, Alexander Wamboldt, and Claude Mellins. 2018. “I Didn’t Want To Be ‘That Girl’”: The Social Risks of Labeling, Telling, and Reporting Sexual Assault. *Sociological Science* 5 (2018), 432–460. <https://doi.org/10.15195/v5.a19>
- [70] Aparup Khatua, Erik Cambria, and Apalak Khatua. 2018. Sounds of Silence Breakers: Exploring Sexual Violence on Twitter. In *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*. 397–400. <https://doi.org/10.1109/ASONAM.2018.8508576>
- [71] Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela J. Wisniewski, and Munmun De Choudhury. 2021. A Human-Centered Systematic Literature Review of Cyberbullying Detection Algorithms. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 325 (oct 2021), 34 pages. <https://doi.org/10.1145/3476066>
- [72] Seunghyun Kim, Afsaneh Razi, Gianluca Stringhini, Pamela J. Wisniewski, and Munmun De Choudhury. 2021. You Don’t Know How I Feel: Insider-Outsider Perspective Gaps in Cyberbullying Risk Detection. *Proceedings of the International AAAI Conference on Web and Social Media* 15, 1 (May 2021), 290–302. <https://doi.org/10.1609/icwsm.v15i1.18061>
- [73] Steffen Krüger and Ane Charlotte Spilde. 2020. Judging books by their covers – Tinder interface, usage and sociocultural implications. *Information, Communication & Society* 23, 10 (2020), 1395–1410. <https://doi.org/10.1080/1369118X.2019.1572771> arXiv:<https://doi.org/10.1080/1369118X.2019.1572771>
- [74] Michelle S. Lam, Mitchell L. Gordon, Danaë Metaxa, Jeffrey T. Hancock, James A. Landay, and Michael S. Bernstein. 2022. End-User Audits: A System Empowering Communities to Lead Large-Scale Investigations of Harmful Algorithmic Behavior. *Proc. ACM Hum.-Comput. Interact.* 6, CSCW2, Article 512 (nov 2022), 34 pages. <https://doi.org/10.1145/3555625>
- [75] Carolyn Lauckner, Natalia Truszczynski, Danielle Lambert, Varsha Kottamasu, Saher Meherally, Anne Marie Schipani-McLaughlin, Erica Taylor, and Nathan Hansen. 2019. “Catfishing,” cyberbullying, and coercion: An exploration of the risks associated with dating app use among rural sexual minority males. *Journal of Gay & Lesbian Mental Health* 23, 3 (2019), 289–306. <https://doi.org/10.1080/19359705.2019.1587729> arXiv:<https://doi.org/10.1080/19359705.2019.1587729>
- [76] Min Kyung Lee and Su Baykal. 2017. Algorithmic Mediation in Group Decisions: Fairness Perceptions of Algorithmically Mediated vs. Discussion-Based Social Division. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Portland, Oregon, USA) (*CSCW ’17*). Association for Computing Machinery, New York, NY, USA, 1035–1048. <https://doi.org/10.1145/2998181.2998230>
- [77] Min Kyung Lee, Ji Tae Kim, and Leah Lizarondo. 2017. A Human-Centered Approach to Algorithmic Services: Considerations for Fair and Motivating Smart Community Service Management That Allocates Donations to Non-Profit Organizations. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) (*CHI ’17*). Association for Computing Machinery, New York, NY, USA, 3365–3376. <https://doi.org/10.1145/3025453.3025884>
- [78] Min Kyung Lee, Daniel Kusbit, Anson Kahng, Ji Tae Kim, Xinran Yuan, Allissa Chan, Daniel See, Ritesh Noothigattu, Siheon Lee, Alexandros Psomas, and Ariel D. Procaccia. 2019. WeBuildAI: Participatory Framework for Algorithmic Governance. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 181 (nov 2019), 35 pages. <https://doi.org/10.1145/3359283>
- [79] Min Kyung Lee, Ishan Nigam, Angie Zhang, Joel Afriyie, Zhizhen Qin, and Sicun Gao. 2021. Participatory Algorithmic Management: Elicitation Methods for Worker Well-Being Models. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) (*AIES ’21*). Association for Computing Machinery, New York, NY, USA, 715–726. <https://doi.org/10.1145/3461702.3462628>

- [80] David Leimstädtner, Peter Sörries, and Claudia Müller-Birn. 2022. Unfolding Values through Systematic Guidance: Conducting a Value-Centered Participatory Workshop for a Patient-Oriented Data Donation. In *Proceedings of Mensch Und Computer 2022* (Darmstadt, Germany) (*MuC '22*). Association for Computing Machinery, New York, NY, USA, 477–482. <https://doi.org/10.1145/3543758.3547560>
- [81] Dennis H Li, Shruta Rawat, Jayson Rhoton, Pallav Patankar, Maria L Ekstrand, BR Rosser, and J Michael Wilkerson. 2017. Harassment and violence among men who have sex with men (MSM) and hijras after reinstatement of India's "Sodomy Law". *Sexuality research and social policy* 14, 3 (2017), 324–330.
- [82] Christian Licoppe, Carole Anne Rivière, and Julien Morel. 2016. Grindr casual hook-ups as interactional achievements. *New Media & Society* 18, 11 (2016), 2540–2558. <https://doi.org/10.1177/1461444815589702> arXiv:<https://doi.org/10.1177/1461444815589702>
- [83] Yingchi Liu, Quanzhi Li, Xiaozhong Liu, Qiong Zhang, and Luo Si. 2019. Sexual Harassment Story Classification and Key Information Identification. In *Proceedings of the 28th ACM International Conference on Information and Knowledge Management* (Beijing, China) (*CIKM '19*). Association for Computing Machinery, New York, NY, USA, 2385–2388. <https://doi.org/10.1145/3357384.3358146>
- [84] Richard MacKinnon. 1997. Virtual rape. *Journal of Computer-Mediated Communication* 2, 4 (1997), JCMC247.
- [85] Wookjae Maeng and Joohnwan Lee. 2022. Designing and Evaluating a Chatbot for Survivors of Image-Based Sexual Abuse. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (*CHI '22*). Association for Computing Machinery, New York, NY, USA, Article 344, 21 pages. <https://doi.org/10.1145/3491102.3517629>
- [86] Julia Mayer and Quentin Jones. 2016. Encount'r: Exploring Passive Context-Awareness for Opportunistic Social Matching. *Proceedings of the 19th ACM Conference on Computer Supported Cooperative Work and Social Computing Companion*, 349–352.
- [87] Weishan Miao and Lik Sam Chan. 2020. Domesticating Gay Apps: An Intersectional Analysis of the Use of Blued Among Chinese Gay Men. *Journal of Computer-Mediated Communication* 26, 1 (12 2020), 38–53. <https://doi.org/10.1093/jcmc/zmaa015> arXiv:<https://academic.oup.com/jcmc/article-pdf/26/1/38/36121062/zmaa015.pdf>
- [88] Miro. 2022. *Miro online whiteboard*. www.miro.com
- [89] Nabila Rezwana Mirza, Shareen Mahmud, Prosonna Hossain Nabila, and Nova Ahmed. 2016. Poster: Protibaadi: An Extended Solution to Deal with Sexual Harassment. In *Proceedings of the 14th Annual International Conference on Mobile Systems, Applications, and Services Companion* (Singapore, Singapore) (*MobiSys '16 Companion*). Association for Computing Machinery, New York, NY, USA, 59. <https://doi.org/10.1145/2938559.2948796>
- [90] Manisha Mohan, Misha Sra, and Chris Schmandt. 2017. Technological interventions to detect, communicate and deter sexual assault. In *Proceedings of the 2017 ACM International Symposium on Wearable Computers*. 126–129.
- [91] Cecily Morrison, Edward Cutrell, Anupama Dhareshwar, Kevin Doherty, Anja Thieme, and Alex Taylor. 2017. Imagining Artificial Intelligence Applications with People with Visual Disabilities Using Tactile Ideation. In *Proceedings of the 19th International ACM SIGACCESS Conference on Computers and Accessibility* (Baltimore, Maryland, USA) (*ASSETS '17*). Association for Computing Machinery, New York, NY, USA, 81–90. <https://doi.org/10.1145/3132525.3132530>
- [92] Michael J Muller and Sarah Kuhn. 1993. Participatory design. *Commun. ACM* 36, 6 (1993), 24–28.
- [93] Fayika Farhat Nova, MD, Rashidujaman Rifat, Pratyasha Saha, Syed Ishtiaque Ahmed, and Shion Guha. 2019. Online Sexual Harassment over Anonymous Social Media in Bangladesh. In *Proceedings of the Tenth International Conference on Information and Communication Technologies and Development* (Ahmedabad, India) (*ICTD '19*). Association for Computing Machinery, New York, NY, USA, Article 1, 12 pages. <https://doi.org/10.1145/3287098.3287107>
- [94] University of Wollongong. 2020. 2020: Ai Research to aid women's safety on public transport - university of wollongong. *UOW* (Aug 2020). <https://www.uow.edu.au/media/2020/ai-research-to-aid-womens-safety-on-public-transport.php>
- [95] World Health Organization et al. 2021. Violence against women prevalence estimates, 2018: global, regional and national prevalence estimates for intimate partner violence against women and global and regional prevalence estimates for non-partner sexual violence against women. (2021).
- [96] Jonathan Petrychyn, Diana C. Parry, and Corey W. Johnson. 2020. Building community, one swipe at a time: hook-up apps and the production of intimate publics between women. *Health Sociology Review* 29, 3 (2020), 249–263. <https://doi.org/10.1080/14461242.2020.1779106> arXiv:<https://doi.org/10.1080/14461242.2020.1779106> PMID: 33411602.
- [97] John R. Porter, Kiley Sobel, Sarah E. Fox, Cynthia L. Bennett, and Julie A. Kientz. 2017. Filtered Out: Disability Disclosure Practices in Online Dating Communities. *Proc. ACM Hum.-Comput. Interact.* 1, CSCW, Article 87 (dec 2017), 13 pages. <https://doi.org/10.1145/3134722>
- [98] Kane Race. 2015. Speculative pragmatism and intimate arrangements: online hook-up devices in gay life. *Culture, Health & Sexuality* 17 (2015), 496 – 511.
- [99] Afsaneh Razi, Ashwaq Alsoubai, Seunghyun Kim, Nurun Naher, Shiza Ali, Gianluca Stringhini, Munmun De Choudhury, and Pamela J. Wisniewski. 2022. Instagram Data Donation: A Case Study on Collecting Ecologically Valid Social

- Media Data for the Purpose of Adolescent Online Risk Detection. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems* (New Orleans, LA, USA) (CHI EA '22). Association for Computing Machinery, New York, NY, USA, Article 39, 9 pages. <https://doi.org/10.1145/3491101.3503569>
- [100] Afsaneh Razi, Seunghyun Kim, Ashwaq Alsoubai, Gianluca Stringhini, Thamar Solorio, Munmun De Choudhury, and Pamela J. Wisniewski. 2021. A Human-Centered Systematic Literature Review of the Computational Approaches for Online Sexual Risk Detection. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 465 (oct 2021), 38 pages. <https://doi.org/10.1145/3479609>
- [101] H. Rosa, N. Pereira, R. Ribeiro, P.C. Ferreira, J.P. Carvalho, S. Oliveira, L. Coheur, P. Paulino, A.M. Veiga Simão, and I. Trancoso. 2019. Automatic cyberbullying detection: A systematic review. *Computers in Human Behavior* 93 (2019), 333–345. <https://doi.org/10.1016/j.chb.2018.12.021>
- [102] Janine Rowse, Caroline Bolt, and Sanjeev Gaya. 2020. Swipe right: The emergence of dating-app facilitated sexual assault. *A descriptive retrospective audit of forensic examination caseload in an Australian metropolitan service* 52 (2020), 1–7. <https://doi.org/10.1007/s12024-019-00201-7> PMID: 32026384.
- [103] Sohini Roy, Abhijit Sharma, and Uma Bhattacharya. 2015. MoveFree: A Ubiquitous System to Provide Women Safety. In *Proceedings of the Third International Symposium on Women in Computing and Informatics* (Kochi, India) (WCI '15). Association for Computing Machinery, New York, NY, USA, 545–552. <https://doi.org/10.1145/2791405.2791415>
- [104] Jennifer D. Rubin, Lindsay Blackwell, and Terri D. Conley. 2020. Fragile Masculinity: Men, Gender, and Online Harassment. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) (CHI '20). Association for Computing Machinery, New York, NY, USA, 1–14. <https://doi.org/10.1145/3313831.3376645>
- [105] Napa Sae-Bae, Xiaoxi Sun, Husrev T. Sencar, and Nasir D. Memon. 2014. Towards automatic detection of child pornography. In *2014 IEEE International Conference on Image Processing (ICIP)*. 5332–5336. <https://doi.org/10.1109/ICIP.2014.7026079>
- [106] Semiu Salawu, Yulan He, and Joanna Lumsden. 2020. Approaches to Automated Detection of Cyberbullying: A Survey. *IEEE Transactions on Affective Computing* 11, 1, 3–24. <https://doi.org/10.1109/TAFFC.2017.2761757>
- [107] Nithya Sambasivan, Amna Batool, Nova Ahmed, Tara Matthews, Kurt Thomas, Laura Sanelly Gaytán-Lugo, David Nemer, Elie Bursztein, Elizabeth Churchill, and Sunny Consolvo. 2019. "They Don't Leave Us Alone Anywhere We Go": Gender and Digital Abuse in South Asia. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland Uk) (CHI '19). Association for Computing Machinery, New York, NY, USA, 1–14. <https://doi.org/10.1145/3290605.3300232>
- [108] Muhammad Yasir Sarosh, Muhammad Abdullah Yousaf, Mair Muteeb Javed, and Suleman Shahid. 2016. MehfoozAurat: Transforming Smart Phones into Women Safety Devices Against Harassment. In *Proceedings of the Eighth International Conference on Information and Communication Technologies and Development* (Ann Arbor, MI, USA) (ICTD '16). Association for Computing Machinery, New York, NY, USA, Article 61, 4 pages. <https://doi.org/10.1145/2909609.2909645>
- [109] Devansh Saxena, Karla Badillo-Urquiola, Pamela J. Wisniewski, and Shion Guha. 2021. A Framework of High-Stakes Algorithmic Decision-Making for the Public Sector Developed through a Case Study of Child-Welfare. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW2, Article 348 (oct 2021), 41 pages. <https://doi.org/10.1145/3476089>
- [110] Devansh Saxena and Shion Guha. 2020. Conducting Participatory Design to Improve Algorithms in Public Services: Lessons and Challenges. In *Conference Companion Publication of the 2020 on Computer Supported Cooperative Work and Social Computing* (Virtual Event, USA) (CSCW '20 Companion). Association for Computing Machinery, New York, NY, USA, 383–388. <https://doi.org/10.1145/3406865.3418331>
- [111] Carol F Scott, Gabriela Marcu, Riana Elyse Anderson, Mark W Newman, and Sarita Schoenebeck. 2023. Trauma-Informed Social Media: Towards Solutions for Reducing and Healing Online Harm. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (CHI '23). Association for Computing Machinery, New York, NY, USA, 1–20. <https://doi.org/10.1145/3544548.3581512>
- [112] Anya Skatova and James Goulding. 2019. Psychology of personal data donation. *PLOS ONE* 14, 11 (11 2019), 1–20. <https://doi.org/10.1371/journal.pone.0224240>
- [113] Peter Snyder, Periwinkle Doerfler, Chris Kanich, and Damon McCoy. 2017. Fifteen Minutes of Unwanted Fame: Detecting and Characterizing Doxing. In *Proceedings of the 2017 Internet Measurement Conference* (London, United Kingdom) (IMC '17). Association for Computing Machinery, New York, NY, USA, 432–444. <https://doi.org/10.1145/3131365.3131385>
- [114] Zahra Stardust, Rosalie Gillett, and Kath Albury. 2022. Surveillance does not equal safety: Police, data and consent on dating apps. *Crime, Media, Culture* 0, 0 (2022), 1741659022111827. <https://doi.org/10.1177/1741659022111827> arXiv:<https://doi.org/10.1177/1741659022111827>
- [115] Denny L. Starks, Tawanna Dillahunt, and Oliver L. Haimson. 2019. Designing Technology to Support Safety for Transgender Women & Non-Binary People of Color. In *Companion Publication of the 2019 on Designing Interactive Systems Conference 2019 Companion* (San Diego, CA, USA) (DIS '19 Companion). Association for Computing Machinery,

- New York, NY, USA, 289–294. <https://doi.org/10.1145/3301019.3323898>
- [116] Elizabeth P. Stedman. 2007. Myspace, but Whose Responsibility - Liability of Social-Networking Websites When Offline Sexual Assault of Minors Follows Online Interaction. *Jeffrey S. Moorad Sports Law Journal* 14 (2007), 363.
- [117] Justyna Stypinska. 2023. AI ageism: a critical roadmap for studying age discrimination and exclusion in digitalized societies. *AI & SOCIETY* 38 (apr 1 2023). Issue 2. <https://doi.org/10.1007/s00146-022-01553-5>
- [118] Ashima Suvarna, Grusha Bhalla, Shailender Kumar, and Ashi Bhardwaj. 2020. Identifying Victim Blaming Language in Discussions about Sexual Assaults on Twitter. In *International Conference on Social Media and Society* (Toronto, ON, Canada) (*SMSociety'20*). Association for Computing Machinery, New York, NY, USA, 156–163. <https://doi.org/10.1145/3400806.3400825>
- [119] Han Tao. 2022. Loving strangers, avoiding risks: Online dating practices and scams among Chinese lesbian (lala) women. *Media, Culture & Society* 44, 6 (2022), 1199–1214. <https://doi.org/10.1177/01634437221088952> arXiv:<https://doi.org/10.1177/01634437221088952>
- [120] Muhammad Uzair Tariq, Afsaneh Razi, Karla A. Badillo-Urquiola, and Pamela J. Wisniewski. 2019. A Review of the Gaps and Opportunities of Nudity and Skin Detection Algorithmic Research for the Purpose of Combating Adolescent Sexting Behaviors. In *Interacción*.
- [121] Rachel Thompson. 2022. Bumble makes cyberflashing detection tool available as open-source code. *Mashable* (Oct 2022). <https://mashable.com/article/bumble-cyberflashing-private-detector-open-source>
- [122] Elisabeth Timmermans and Cédric Courtois. 2018. From swiping to casual sex and/or committed relationships: Exploring the experiences of Tinder users. *The Information Society* 34, 2 (2018), 59–70. <https://doi.org/10.1080/01972243.2017.1414093> arXiv:<https://doi.org/10.1080/01972243.2017.1414093>
- [123] Elisabeth Timmermans and Elien De Caluwé. 2017. Development and validation of the Tinder Motives Scale (TMS). *Computers in Human Behavior* 70 (2017), 341–350. <https://doi.org/10.1016/j.chb.2017.01.028>
- [124] Julie L. Valentine, Leslie W. Miles, Kristen Mella Hamblin, and Aubrey Worthen Gibbons. 2022. Dating App Facilitated Sexual Assault: A Retrospective Review of Sexual Assault Medical Forensic Examination Charts. *Journal of Interpersonal Violence* 0, 0 (2022), 08862605221130390. <https://doi.org/10.1177/08862605221130390> arXiv:<https://doi.org/10.1177/08862605221130390> PMID: 36310506.
- [125] Kristin Veel and Nanna Bonde Thylstrup. 2018. Geolocating the stranger: the mapping of uncertainty as a configuration of matching and warranting techniques in dating apps. *Journal of Aesthetics & Culture* 10, 3 (2018), 43–52. <https://doi.org/10.1080/20004214.2017.1422924> arXiv:<https://doi.org/10.1080/20004214.2017.1422924>
- [126] George Veletsianos, Shandell Houlden, Jaigris Hodson, and Chandell Gosse. 2018. Women scholars' experiences with online harassment and abuse: Self-protection, resistance, acceptance, and self-blame. *New Media & Society* 20, 12 (2018), 4689–4708. <https://doi.org/10.1177/1461444818781324> arXiv:<https://doi.org/10.1177/1461444818781324>
- [127] Emily A Vogels. 2020. 10 facts about Americans and online dating. <https://www.pewresearch.org/fact-tank/2020/02/06/10-facts-about-americans-and-online-dating/>
- [128] Ashley Marie Walker and Michael A DeVito. 2020. "More gay'fits in better": Intracommunity Power Dynamics and Harms in Online LGBTQ+ Spaces. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.
- [129] Hao Wang, Congxing Cai, Andrew Philpot, Mark Latonero, Eduard H. Hovy, and Donald Metzler. 2012. Data Integration from Open Internet Sources to Combat Sex Trafficking of Minors. In *Proceedings of the 13th Annual International Conference on Digital Government Research* (College Park, Maryland, USA) (*dg.o'12*). Association for Computing Machinery, New York, NY, USA, 246–252. <https://doi.org/10.1145/2307729.2307769>
- [130] Mark Warner, Andreas Gutmann, M. Angela Sasse, and Ann Blandford. 2018. Privacy Unraveling Around Explicit HIV Status Disclosure Fields in the Online Geosocial Hookup App Grindr. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 181 (nov 2018), 22 pages. <https://doi.org/10.1145/3274450>
- [131] Mark Warner, Juan F. Maestre, Jo Gibbs, Chia-Fang Chung, and Ann Blandford. 2019. Signal Appropriation of Explicit HIV Status Disclosure Fields in Sex-Social Apps Used by Gay and Bisexual Men. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (Glasgow, Scotland UK) (*CHI '19*). Association for Computing Machinery, New York, NY, USA, 1–15. <https://doi.org/10.1145/3290605.3300922>
- [132] Meredith Whittaker, Meryl Alper, Cynthia L Bennett, Sara Hendren, Liz Kazianus, Mara Mills, et al. 2019. Disability, bias, and AI. (2019).
- [133] Heather Wolbers, Hayley Boxall, Cameron Long, and Adam Gunnoo. 2022. *Sexual harassment, aggression and violence victimisation among mobile dating app and website users in Australia*. <https://doi.org/10.52922/rr78740>
- [134] Christine T. Wolf, Haiyi Zhu, Julia Bullard, Min Kyung Lee, and Jed R. Brubaker. 2018. The Changing Contours of "Participation" in Data-Driven, Algorithmic Ecosystems: Challenges, Tactics, and an Agenda. In *Companion of the 2018 ACM Conference on Computer Supported Cooperative Work and Social Computing* (Jersey City, NJ, USA) (*CSCW '18*). Association for Computing Machinery, New York, NY, USA, 377–384. <https://doi.org/10.1145/3272973.3273005>

- [135] Peng Yan, Linjing Li, Weiyun Chen, and Daniel Zeng. 2019. Quantum-Inspired Density Matrix Encoder for Sexual Harassment Personal Stories Classification. In *2019 IEEE International Conference on Intelligence and Security Informatics (ISI)*. 218–220. <https://doi.org/10.1109/ISI.2019.8823281>
- [136] Chaeyoon Yoo and Paul Dourish. 2021. Anshimi: Women’s Perceptions of Safety Data and the Efficacy of a Safety Application in Seoul. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 147 (apr 2021), 21 pages. <https://doi.org/10.1145/3449221>
- [137] Wenqi Zheng, Emma Walquist, Isha Datey, Xiangyu Zhou, Kelly Berishaj, Melissa Mcdonald, Michele Parkhill, Dongxiao Zhu, and Douglas Zytko. 2024. “It’s Not What We Were Trying to Get At, but I Think Maybe It Should Be”: Learning How to Do Trauma-Informed Design with a Data Donation Platform for Online Dating Sexual Violence. In *Proceedings of the CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI ’24)*. Association for Computing Machinery, New York, NY, USA, Article 743, 15 pages. <https://doi.org/10.1145/3613904.3642045>
- [138] Haiyi Zhu, Bowen Yu, Aaron Halfaker, and Loren Terveen. 2018. Value-Sensitive Algorithm Design: Method, Case Study, and Lessons. *Proc. ACM Hum.-Comput. Interact.* 2, CSCW, Article 194 (nov 2018), 23 pages. <https://doi.org/10.1145/3274463>
- [139] Caleb Ziems, Ymir Vigfusson, and Fred Morstatter. 2020. Aggressive, Repetitive, Intentional, Visible, and Imbalanced: Refining Representations for Cyberbullying Classification. *Proceedings of the International AAAI Conference on Web and Social Media* 14, 1 (May 2020), 808–819. <https://doi.org/10.1609/icwsm.v14i1.7345>
- [140] Douglas Zytko, Nicholas Furlo, Bailey Carlin, and Matthew Archer. 2021. Computer-Mediated Consent to Sex: The Context of Tinder. *Proc. ACM Hum.-Comput. Interact.* 5, CSCW1, Article 189 (apr 2021), 26 pages. <https://doi.org/10.1145/3449288>
- [141] Douglas Zytko, Pamela J. Wisniewski, Shion Guha, Eric P. S. Baumer, and Min Kyung Lee. 2022. Participatory Design of AI Systems: Opportunities and Challenges Across Diverse Users, Relationships, and Application Domains. In *Extended Abstracts of the 2022 CHI Conference on Human Factors in Computing Systems (New Orleans, LA, USA) (CHI EA ’22)*. Association for Computing Machinery, New York, NY, USA, Article 154, 4 pages. <https://doi.org/10.1145/3491101.3516506>
- [142] Douglas Zytko, Nicholas Mullins, Shelnesha Taylor, and Richard H. Holler. 2022. Dating Apps Are Used for More Than Dating: How Users Disclose and Detect (Non-)Sexual Interest in People-Nearby Applications. *Proc. ACM Hum.-Comput. Interact.* 6, GROUP, Article 30 (jan 2022), 14 pages. <https://doi.org/10.1145/3492849>
- [143] Douglas Zytko, Victor Regalado, Nicholas Furlo, Sukeshini A. Grandhi, and Quentin Jones. 2020. Supporting Women in Online Dating with a Messaging Interface That Improves Their Face-to-Face Meeting Decisions. *Proc. ACM Hum.-Comput. Interact.* 4, CSCW2, Article 137 (oct 2020), 30 pages. <https://doi.org/10.1145/3415208>

A Additional Methodological Details

A.1 Participant Demographic Details

Table 10. Demographic details of interview participants.

Participant	Ethnicity	Age	Apps used	Goals for app-use	Risk or harm experienced
P1	White, Hispanic	30	OkCupid	Dating	Unwanted touching
P2	White	38	Facebook Dating, match.com, eHarmony, Christian Mingle, Zoosk	Dating	Did not trust meeting partner during date
P3	Middle Eastern or North African	29	Muzmatch, Salams, Mustinder, Hinge	Dating, group events	Racist abuse
P4	White	25	Instagram	Friends, dating	Unwanted touching
P5	Asian	32	Meetup, Facebook, Instagram	Friends, professional networking	N/A
P6	Asian	33	Tinder, Bumble, OkCupid, Hinge, Coffee Meets Bagel	Dating	Meeting partner left her alone in his house

P7	Asian	24	Meetup, Discord	Friends, group events	N/A
P8	White	22	Tinder, Bumble, OkCupid	Dating	Drugged without consent
P9	Black	20	Bumble, Hinge	Dating	N/A
P10	White	41	Tinder, OkCupid	Friends, dating	N/A
P11	Asian	28	Instagram	Group events, friends	N/A
P12	Black	26	Tinder	Dating	Attempted scam
P13	Native American	18	Snapchat, Instagram, Patio	No specific goal	N/A
P14	White	30	Tinder, Bumble, OkCupid	No specific goal, friends	Misogyny, concern for safety
P15	White	25	Tinder, Bumble	Dating	Unwanted sexual advances by drunk partner
P16	Black	24	Tinder	Dating	Partner lied about important details
P17	Black	25	Tinder	Friends, dating	Partner lied about important details
P18	White, Black	28	OkCupid	Dating	Partner lied about important details
P19	Black	27	Tinder	Dating	Partner hid severe alcoholism
P20	Black	35	Tinder	Dating	Partner demanded sex

A.2 Model Template Used in Model Building Activity and an Example of Populating it with a Feature

Table 11. The risk detection model template provided to each participant. Each suggested feature includes various possible feature states the participant is likely to encounter and the weights based on their preferred risk scale that they assign to each feature state signifying how important or unimportant that state is in their risk assessment

Feature 1	Feature State A	Feature State B	Feature State C
Weights			

Table 12. An example of populating the model template with a feature for "criminal record of other person" from P4's model. The possible states of the feature are indicated in each column (i.e., the person depicted in the profile could either have no criminal record, a misdemeanor charge, or a felony charge). The "weights" row indicates how each possible state would contribute to the numerical assessment of risk. A higher weight means more risk. For instance, having a felony charge would add 10 points to the person's overall risk score, whereas having no criminal record would only add 2 points.

Criminal Record Of Other Person	No Criminal Record	Misdemeanor Charge	Felony Charge
Weights	2	6	10

B Descriptive Account of Participants' Risk Detection Models

In this section we provide a descriptive account of the risk detection models created by participants to lend context to the qualitative findings in the findings. Participants proposed 45 unique features


for their risk detection models. On average models contained 10.8 ± 1.86 features. Table 13 compiles the models from all 20 participants. It lists the 45 features on the y axis and participants along the x axis (P12 is listed twice because they produced separate models for individual meeting partners and group activities). Features are sorted in descending order by the number of models the feature was present in (most common features near the top). "Feature weight range" refers to the minimum and maximum weights applied to features in a participant's model. The presence of a feature in a participant's model is indicated by the respective cell containing a number and cell color. The number corresponds to the highest possible weight for the feature in that participant's model; a higher weight means higher risk. Since participants had different feature weight ranges (up to 10, up to 5, and up to 3), we converted them all to a standard scale of 1-6 and used a color gradient to visually indicate a feature's maximum risk relative to the maximum possible (6) risk in that participant's model: (lowest risk) 1  6 (highest risk). The darker the color (closest to dark red) the higher the maximum risk of the feature, and the brighter the color (closest to yellow/green), lower the maximum risk of the feature. In other words, bright colored features have the potential to impact risk the most. The maximum relative risk for each feature can also be determined without cell colors by cross-referencing the feature's maximum risk score with the participant's feature weight range. A profile's total risk score is a summation of the scores for every feature, with the exception of group-based activity features (italicized) which would not be applied to assessing risk of individual meeting partners. Refer to Table 14 in the appendix for definitions of every feature and example feature states.

Table 13. Visualization of all participants' risk detection models

Participant	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P12 (g)	P13	P14	P15	P16	P17	P18	P19	P20
Feature weight range	1-10	1-10	1-10	0-10	1-10	1-3	0-5	0-10	1-10	1-5	0-10	1-10	1-10	1-10	1-5	0-5	1-5	1-10	1-10	1-5	1-3
No. of meeting partners	5.4	5.4	6	4.5	6	6		4.3	3.8	2.3		4.9	4.9	3.8	6	4	6	5.4	3.8	4.8	6
Location familiarity	5.4	4.3	6			6	4	4.5		3.5	5	5.4	4.9	6	4.8	5	6	3.8	4.9	4.8	3.5
Meeting time	6	4.9	4.6	5.8	5.2	6			3.8			4.9	4.9		3.5	6	6	5.4	6	6	6
Bystander presence	6	5.4	6	6	6			3.8	4.9		6	4.3	4.3		5.4		6	4.9	5.4	4.8	6
Location reputation		5.4		5	5.7				4.9	5.4	5.8	6	6		3.5		6		4.9	6	5
Gender	4.9	4.6		6		6		4.5		4.8				4.9	6	4	6	4.9		4.8	
WiFi/phone reception				6			6	5.5	6	6		6	6	6	6	6	6				6
Criminal record		5.7		6	6						6	5.4		6	3.5	5	6		4.3	3.5	
Familiarity with meeting partner(s)	6	5.4		4	4.6	4	3.5							4.3		5	6				6
Security/police proximity				5.5			4	3.8				5.4	4.3	5.2		3	6	3.8			
Age difference		4.3	4.6	6				6	6					3.8						3.5	
Profile completeness	6				6	6		6		4.8	5.8										6
Location publicness			6			6			5.4						6	6		4.3			

Proximity to home	4.9		4.8	6	4	2.9			4.6						
Presence of friends		4.6					3.5	4.5				4.9		6	
Presence of alcohol			5.8	6						3.5	6			3.5	
Reputation of meeting partner(s)	6				5	5.4	5.8		4.9						
Communication before meeting					6		3.5	5.3					4.3		
Able to view profile(s) before meeting								4.9	4.9				4.3	4.8	3.5
Familiarity b/w meeting partners			5.4		5								5.5		
Easy to enter/leave location				6										4.8	6
Proximity to vehicle/transport				6	6	5									
Location type		3.8										6		3.2	
Mutual connections		4.9			6			3.8							
Spiritual values		4.3	6												
Education level		5.4	6												
Day of week for meeting			6	4.3											
Political affiliation							5.4					5.4			
Location population density							3.5					3.5			
Recreational drug use						5							3.5		
Face-to-face meeting history							3.8	2.5							
Criminal record in group								5							
Meeting end time													5.3		
Group activity excitement	6														
Photo of group activity				6											
Shared interests					6										
User-determined red flags										6					
Mental health													4.8		
Presence of activity host								6							
Location cleanliness							4.9								
Marital status		4.9													

Cultural background	4.9																			
Personality			4.6																	
Shared student status										4.3		3								
Passive observation of partner									2.7											

C Risk Detection Model Features and Definitions

Table 14. Model features, their definitions, and example feature states

Feature	Definition	Example Feature States
No. of meeting partners	Number of people that user is meeting face-to-face	One-on-one, <4 participants, >4 participants (P1)
Location familiarity	Prior familiarity with meeting location	Location is very familiar, occasionally visited, not been to before (P2)
Meeting time	Time of face-to-face meeting/activity	Morning, afternoon, evening (after 6:30-7 PM) (P3)
Bystander presence	Presence of others at meeting location to seek help from	0, 1-2 surrounding people, 3-10 surrounding people, 10+ (P4)
Location reputation	Meeting location crime rate, reviews	No known crime, crime occurs regularly (monthly), frequent crime (weekly) (P5)
WiFi/phone reception	Availability of WiFi or cell-phone reception at meeting location	No connectivity, limited connectivity, full connectivity (P7)
Familiarity with meeting partner(s)	Does user already know the person/the people they are meeting?	Know the user (1-on-1), know at least one (group) attendee, do not know anyone (P16)
Gender	Gender of one-on-one meeting partner or gender distribution of group activity	All male, mix of genders, all female (P6)
Criminal record	Does one-on-one meeting partner have a criminal record?	No Criminal history/only petty offenses (jaywalking/pranks/shoplifting), violent conviction, convictions viewed positively (civil rights) (P13)
Age difference	Age difference with meeting partner(s)	Same age (2 years younger or 4 years older), older, younger, minor (P8)
Security/police proximity	How close are police or professional security personnel?	None, at least one at location, 2-3 present (P15)
Profile completeness	How complete is a user/social opportunity profile on the app?	No text content, full text description (P1)
Proximity to home	How close does the user live to the meeting location home?	<2 miles, 2-10 miles, >10 miles (P6)
<i>Presence of friends</i>	Is the user attending a group activity with friends?	Is Alone, with an acquaintance, with a good friend (P10)
Presence of alcohol	Is alcohol present at meeting location?	Alcohol present, alcohol is not present (P14)

Reputation of meeting partners	Status of user/social opportunity reviews (necessitates review functionality in-app)	Reviews are 100% positive, mixed review scores, 100% negative (P7)
Location publicness	Is the meeting location a public place?	Private residence, business establishment, public space (e.g., park) (P17)
<i>Criminal record in group</i>	Do any attendees of group activity have a criminal record?	No one has a criminal record, at least one person has a criminal record (P11)
Communication before meeting	Can the participant communicate with potential meeting partner(s) before face-to-face encounter?	They can chat beforehand, unable to communicate beforehand (P8)
<i>Familiarity between meeting partners</i>	Do the other attendees already know each other?	Other attendees are friends, relatives, all strangers to each other (P4)
Easy To enter/leave meeting location	Capacity to quickly leave an unsafe situation	Able to leave easily/at any time, difficult to leave one-one-one meeting, difficult to leave group activity (P20)
Proximity to vehicle/transport	How far is the user's car or other preferred transportation?	Near, kind of far (within 10 minutes), far away (P8)
Location type	What type of location is the meeting at?	Restaurant, bar, public park, library (P2)
Spiritual values	Does the meeting partner(s) share user's spiritual values?	Muslim, not Muslim (P3)
Mutual connections	Meeting partner in same social circle as user	Person is connected to people who share values, person is connected to people who do not share values, no mutual connections (P2)
Education level	Education level of meeting partner	Less than Bachelor's degree, Bachelor's degree, Master's degree or PhD (P3)
Day of week for meeting	What day of the week is the meeting scheduled for?	Weekday, weekend (P5)
Political affiliation	Does meeting partner(s) have contrasting political views?	Divisive content in profile, No divisive profile content (P10)
Location population density (rurality)	How densely populated is the area surrounding the meeting location?	Area is remote (e.g., hiking trail), area is rural, area is urban/metropolitan (P10)
Meeting end time	What time is the meeting expected to end?	During daytime, evening (6-8 PM), late night (past 8 PM) (P16)
<i>Group activity excitement</i>	Is the user familiar with and excited about what normally happens at the respective activity?	Not exciting, sort of exciting, especially exciting (P1)
<i>Photo of group activity</i>	Is there a live photo of an ongoing activity uploaded on the app? (requires new app functionality)	Picture available and looks safe, looks unsafe, photo is generic location image (P5)
Shared interests	Does the user/attendee share interests with participant?	Share interests, no shared interests (P7)
User-determined red flags	Does user manually identify indicators of risk in profile (requires direct user input to AI)?	2+ red flags, 1 red flag, no red flags (P13)

Mental health	Does meeting partner(s) have mental health issues that could result in erratic/dangerous behavior?	1+ meeting partner has a mental illness, none are known to have a mental illness (P16)
<i>Presence of activity host</i>	Does group activity have host to manage/monitor safety?	There is host I can communicate with, no host (P11)
Location cleanliness	Location is generally clean and well-maintained	Area is well maintained, somewhat maintained, not clean (litter, overgrown) (P9)
Marital status	Marital status of meeting partner	Single, married, married with children (P2)
Cultural background	Is partner associated with a culture that has regressive views about women?	Partner is from a culture that does not treat women equally, person is from a culture where women are treated equally (P2)
Recreational drug use	Is the meeting location/type associated with drug use?	Drug use is present, drug use is not present but alcohol is, no drug presence (P8)
Personality	Personality traits meeting partner(s) (anticipated data sources described in 4.4)	Introverted/socially awkward, Extroverted, mixed personalities (group) (P5)
Shared student status	Is any meeting partner a university student like user (only selected by participants who were also students)?	No student(s), at least one student, there more than one student (group activity) (P13)
Face-to-face meeting history	How many other people a partner has previously met through the app?	Long history of prior face-to-face encounters, 1-5 prior encounters, no face-to-face encounters (P11)
Passive observation of potential meeting partner	Is the user able to physically observe a potential meeting partner right before direct face-to-face interaction?	Able to passively observe the partner (e.g., through a glass window from outside a restaurant), unable to observe partner (P9)

Received July 2023; revised April 2024; accepted July 2024