"Just Like, Risking Your Life Here": Envisioning AI As a Supportive Agent to Bolster Women's Strategies for Assessing Risk of Harm in Online Dating

Isha A. Datey, Douglas Zytko

Oakland University

Abstract

The upsurge of online dating usage has opened additional avenues for infliction of harm, especially against women. This study explores how women's interactions with users and AI contributes and can contribute to respectively ameliorating decision making when choosing whether to initiate communication and whether to convert online communication to offline based on the degree of risk of harm may pose. We evaluate how women envision AI supporting them in assessing risk of harm in online dating, and how this risk assessment AI could influence their current communicative practices. The research, thus, presents an interview study in which twenty cisgender and transgender women participated in focus groups and one-on-one interview sessions, later open coded for qualitative analysis. Women produced AI models augmenting their strategies for assessing risk of harm with communicating through online modalities (textual and otherwise) or meeting given users offline. Contributions include conceptualization of a risk assessment AI that works alongside women instead of replacing their role in risk assessment completely, anticipated features to inform the risk assessment AI model are not only vast but also interdependent in the decision-making process, demonstrated need for user-personalization of risk assessment AI models, and provides methodological suggestions for future work in participatory design of AI.

Keywords: HCI, Women, Online Dating, Empirical Study, Interviews, Artificial Intelligence, Technology, Risk, Harm

“Just Like, Risking Your Life Here”: Envisioning AI As a Supportive Agent to Bolster Women’s Strategies for Assessing Risk of Harm in Online Dating

Up to 800 million women have been victims of sexual or physical violence at least once in their lifetime (World Health Organization, 2021). In response, research has sought to understand the role of computer-mediated communication in gender-based violence (MacKinnon, 1997, Henry & Powell, 2015, Douglass et. Al., 2018, DeKeseredy, 2020, Rubin et. Al., 2020).

Today, Artificial Intelligence (AI)-based risk detection technologies present promising potential to automatically detect risky content for harm mitigation. Among ai-based initiatives, called Modus Operandi Safely (i.e., MOSafely), Alsoubai et. al. (2022) built a youth-centered, web-based risk detection dashboard called, *MOSafely, “Is that Sus?”* which leverages machine learning algorithms developed to detect risks within youth online interactions and provide them the ability to give feedback on the AI suspected risks. In a similar bid to protect a vulnerable demographic from computer facilitated harms, we attempt to design an AI system to address the risk of harm presented by online dating apps, a prominent context of violence against women (UK National Crime Agency. 2016).

An important AI proposal in research, for women specifically, has been to aid female victims of dating violence assess the danger in their abusive relationship, set priorities for safety, and develop a personalized safety plan (Lindsay, et. al, 2013). While other work has involved using AI chatbots to provide informational and mental support to victims of various abuses and harms, like victims of image based sexual abuse (for example revenge porn) (Maeng & Lee., 2022), for detection of nude images in direct messages given the disproportionate cyberflashing victimization of women (Vogels, 2020), and detecting advertisements about trafficked women and children online (Wang et. Al., 2012). The latter can be categorized as risk detection AI, which is a use case of AI for safety where different branches of Artificial intelligence such as machine learning and Natural language processing are utilized to detect harm online (Razi et. Al., 2021). But not necessarily always in a way that assists in preventing harm, as Risk Detection AI may imply that harm needs to first occur for it to be detected. We therefore reword risk detection AI as risk assessment AI because its purpose is not to just detect harm but also act as an evaluator in assessing the risk of harm in online as well as first-time in-person encounters.

Online dating apps serve as an important context for the implementation of risk assessment AI because of the potential of facing risks not just online such as facing cyber flashing (Vogels, 2020), sexual harassment and violence (Powell & Henry, 2019, Rowse, Bolt, & Gaya, 2020, Shapiro et. Al., 2017, Thompson, 2018., National Crime Agency, 2016, Lauckner et. al., 2018), privacy breaches (Lutz & Ranzini, 2017, Brandtzaeg et. Al., 2019, Ghorayshi & Ray, 2018, Blackwell et. Al., 2015), racism (Callander, Holt, & Newman, 2016, Lauckner et. al., 2019, Miao & Chan, 2021), and deception through catfishing (Lauckner et. al., 2019), but the risk of these translating offline as well, such as these leading to traumatization, lgbtq+ users being outed and abused (Li et al., 2017), being financially swindled, in-person rape, injury, spread of HIV/STI cases because of risky sexual practices (Handel et. al. 2012) and so on respectively.

Centelles, et. al. (2021) reported in their study that location-based dating apps, that is dating apps that match people who are near each other (Ma, Sun, and Naaman, 2017), may be especially dangerous environments for women even when seemingly minimizing exposure on their profiles or engaging in harm reduction strategies online (Centelles, Powers, & Moule, 2021). These strategies typically revolve around uncertainty reduction and impression formation which, in this context, refers to reducing uncertainty surrounding other people online based on cues presented through different means (for example in dating apps, text in dating profiles, textual chat conversations, and profile images). This enables a user to understand the person enough to form expectations based on the information offered and strategize accordingly to mitigate any expected harm (Knobloch, 2015). There is a plethora of research that has attempted to capture users' online uncertainty reduction strategies and impression formation (Gibbs et. Al., 2011) and the struggles they face in these to protect themselves from the harm when using online dating apps (Centelles, Powers, & Moule, 2021). Understanding these takes us a step closer to designing solutions that better align the design of risk assessment AI with users' needs.

AI as a design implementation for safety in dating apps has only been seen so far for scam detection and chat moderation in dating apps even though AI has long been a part of its functioning for user discovery (Dash Technologies Inc, 2022). Few new prototypes and design implementations have emerged in the past few years and focus particularly on protecting users from an aspect of online sexual violation, such as "Private detector AI" in Bumble and Badoo for nude image detection in chats (Thompson, 2022), and content moderation AI in its beta-testing stage in Tinder that screens and flags potentially offensive messages and allows users to report the offender based on that (Rivero, 2021). We attempt to reimagine dating apps, through a participatory design approach (Zhang & Wakkary, 2014), where the user is empowered to assess the risk of harm inflicted through online modalities, such as cyber flashing as well as the possibility of offline harms such as rape and injury. Towards this, we involve one of the most vulnerable demographics, women in this context, in co-designing risk detection AI in online dating.

Towards ensuring that the risk awareness AI would meet the needs and expectations of women, this paper presents findings from an interview study with 20 cisgender and transgender women in North America about their experiences with online dating and co-creating personal risk assessment AI models for online dating apps. The study outlines how an AI model for risk assessment that works in partnership with women to increase their awareness of risk may be designed:

RQ1. *How do women conceptualize AI supporting them in assessing the risk of harm in online dating apps?*

RQ2. *How can participants' personalized AI models for assessing the risk of harm contribute to the understanding of how a risk assessment AI should be designed in online dating apps?*

RQ3. *What implications do these personal models have on a possible real-life implementation of such a risk assessment AI?*

Empirical contributions include:

1. Conceptualization of risk assessment AI that works alongside women to augment their existing strategies of assessing risk of harm by acting as a supportive, information gathering entity as opposed to completely replacing the human agent in the risk assessment process.
2. Anticipated features to inform the risk assessment AI model are not only vast but also interdependent in the decision-making process. These features inform three themes that contribute to the risk of harm: Women's ability to seek help during a harmful encounter in-person, their ability to eject physically from a situation, and to gauge the likelihood of users' inflicting harm online and/or offline.

3. Demonstrated need for user-personalization of risk assessment AI models. The wide variety of features, feature weightages, and the differences in how similar features are weighed by different participants in their personal models during the individual model building study suggests the need to retain subjectivity and variability in any real-life implementations of this risk assessment AI.
4. Provides methodological suggestions for future work in participatory design of AI. Co-creating AI models with participants can help build participants' understanding and interpretability of an AI's output. This in turn is important to develop an appropriate degree of participants' trust towards the AI to support optimal human-AI performance.

Background

Disparity in Gender Based Violence and HCI

Violence against women is a constant threat with grave physical and mental implications (World Health Organization, 2021) and as HCI researchers, it is imperative we continue addressing computer-mediation of harms considering the swiftly advancing digital age, especially solutions that are preventative as opposed to just-in-time interventions, which have questionable utility (Karusala & Kumar, 2017), or post-harm support (Andalibi, et. al., 2016).

Up to 800 million women have been victims of sexual or physical violence at least once in their lifetime. Women have disproportionately been the victims of violence, and this is a global issue (World Health Organization, 2021). Research has repetitively confirmed gender disparity in sexual and domestic violence (Meyersfeld, 2012, Watts & Zimmerman, 2002, Smith et. Al., 2016.) In response, HCI researchers have been seeking to understand the role of computer-mediated communication in gender-based violence (MacKinnon, 1997, Henry & Powell, 2015, Douglass et. Al., 2018, DeKeseredy, 2020, Rubin, Blackwell, & Conley, 2020), and especially how technology and its affordances have been playing a role in assisting perpetuation of harm or exacerbating negative attitudes against women (DeKeseredy, 2020).

How technology affordances exacerbate harm can be seen in the way violent, degrading heterosexual porn is freely distributed online and influences men to replicate similar behaviors against women, especially spouses and partners, in real life (DeKeseredy, 2020). Domestic abusers have also been known to exploit technology such as spyware and cyber-attacks to intimidate, threaten, monitor, impersonate, harass, or otherwise harm their victims (Freed et. Al., 2018) for the purpose of causing emotional, physical and financial harm by tracking their location and monitoring communications (Chatterjee et al., 2018).

Gender Disparities in abuse and violence are reportedly repeatedly replicated in online spaces as well (Rubin, Blackwell, & Conley, 2020, Sambasivan et. Al., 2019, Veletsianos et. Al., 2018, MacKinnon, 1997). Sexism, for instance, is rampant in the online video gaming spaces (Fox & Tang, 2014), South Asian women constantly face cyberstalking, impersonation, and personal content leakages online (Sambasivan et. Al., 2019), women scholars experience abuse (Veletsianos et. Al. 2018), and not even virtual reality spaces are safe for women (MacKinnon, 1997).

Concerns have been aired in literature about voluntary and involuntary sexual acts being captured digitally and spread to others without consent from the victim such as sex tapes or sex attack films (Medew, 2007) and revenge porn where intimate nude images are spread to others without partners’ consent, not always motivated by revenge (Henry et. al., 2018, Maeng & Lee, 2022, Citron & Franks, 2014). New technologies have also enabled people with no knowledge of AI and machine learning to spread AI-generated deepfake, revenge porn (Maeng & Lee, 2022) in the absence of real imagery.

Popular technologies such as social media have been utilized to sexually exploit people by luring or forcing minors and adults into sex trafficking. (Razi et. Al., 2021, Wang et. Al., 2012).

Technology for Safety and AI

HCI researchers have been discussing solutions to combat these harms as well as support victims of harm for years now and there has been a strong focus in utilizing AI as a design material for safety implementations, most commonly AI for risk detection.

So far, technological solutions in mitigating harm include preventative measures, just-in-time interventions, and post ham support.

Post harm support technology, though important in mitigating further harm such as more abuse (Maeng & Lee, 2022) or adverse mental health effects (Medew, 2007), do not contribute immensely to preventing or assessing harm in the first place. Helping female survivors of dating violence assess the danger in their abusive relationship, set priorities for safety, and develop a personalized safety plan (Lindsay et. Al., 2013), help women reframe sexual abuse experiences (Dimond et. Al., 2013) or help image based sexual abuse survivors seek information and support in the aftermath (Maeng & Lee, 2022) are just some examples of such post-harm support technology proposed in research.

On the other hand, just-in-time interventions primarily include panic button implementations, which have met criticism for this "just-in-time" nature (Karusala & Kumar, 2017) and can be seen incorporated into infrastructural implementations like Blue Light Emergency phones on many college campuses in the US (Ellcessor, 2019) or have even started pervading policies, like in India (Karusala & Kumar, 2017). It is worth mentioning that most solutions proposed in this category do not always address technology-facilitated harms specifically, but the overarching risk of harm of an in-person encounter, expected and otherwise. Proposed solutions of this kind are wide ranging from Protibadi in Bangladesh (Mirza, et. al., 2016, Ahmed, et. Al., 2014), Naari in India (Veturekar et. Al., 2022), and U-Signal and MoveFree for trans women in the United States (Roy et. Al., 2015, Starks, Dillahunt, & Haimson, 2019) as well as physical, rapist-deterrent wearable technology (Mohan, Sra, & Schmandt, 2017) that emits odor to deter attackers and, a common theme in these solutions, activates alarms to contact help.

Finally, preventative solutions have been discussed globally, prominently in pedestrian context (Blom et. Al., 2010), such as mobile application prototypes that enable vulnerable demographics like women to assess danger by helping them find safe passages to locations (Ali et. Al., 2015), such as MehfoozAurat in Pakistan (Sarosh et. Al., 2016.) and Anshimi in South Korea (Yoo & Dourish, 2021.) Recent research has also involved analyzing and calling for design of social platforms that better support user practices of exchanging interpersonal consent (Im et. Al., 2021) as they are increasingly being used to mediate consent practices (Zytko et. Al., 2021).

However, there has been a steady rise in literature utilizing AI for similar harm mitigation purposes and has especially been prevalent in sexual risk detection, content moderation (Razi et. Al., 2021, Tariq et. Al., 2019), and scam detection (Yurtseven et. Al., 2021, Al-Rousan et. Al., 2020). Content moderation AI not only makes moderating offensive or problematic content on, for instance, social media, scalable and fast but also prevents traumatizing human moderators (Gillespie, 2020)). Additionally, detecting Sexual Grooming (Lorenzo-Dus, Kinzel, & Cristofaro, 2020) on social platforms, according to a literature review conducted by Afsanah Razi et. Al (2021)., is the most prominent focus of risk detection AI and algorithms followed by Sexual harassment/abuse and Sexual trafficking.

However, such AI for risk detection could be considered preventative, in that it prevents attackers from committing further harm by helping in their detection online, such as the case with detecting scammer profiles from social media platforms who attempt to use celebrity pictures for catfishing (Al-Rousan et. Al., 2020). Or it could be considered post-harm support, where it detects, for example, advertisements of children and women being trafficked online (Wang et. Al., 2012), and is therefore detecting risks and harms while or after they have occurred (in this case, traumatization and trafficking).

As such, for this study, we reword Risk detection AI as Risk assessment AI as the former may imply that harm needs to first occur for it to be detected while the latter attempts to assess the possibility of risk, thus creating more awareness in relation to the risk of harm that a user may face.

Risk detection AI in Online Dating Apps

An important context that requires further exploration of the utility of AI in automating risk assessment is online dating applications. As one of the most prominent contexts of online and technology facilitated abuse, the current state of dating apps (UK National Crime Agency. 2016) puts more and more people at risk of harm as the user base expands rapidly (Smith & Anderson, 2015). Therefore, it is imperative to continue researching solutions that address the risk of harm in online dating, especially utilizing AI as a design material to develop more efficient solutions (Stamova & Draganov, 2020).

Risks of harm in online dating

Risk of harm due to other application users include the risk of cyberflashing (Vogels, 2020), catfishing, deception, (Lauckner et. al., 2018, Toma, Hancock, & Ellison, 2008), Scams (Albury et. Al., 2021, Tao, 2022)), discrimination (Lauckner et. al., 2018), bullying (Birnholtz et. Al., 2014), racism (Holt & Newman, 2016), sexual coercion and harassment (Lauckner et. al., 2018), public health risks like rise in cases of HIV/STI in gay dating apps (Warner et. Al., 2018., Warner et. Al., 2019), and risk of privacy breaches by people familiar and otherwise (Lutz & Ranzini 2017) leading to physical and emotional harms such as stalking, ostracization or shaming and even prosecution in the case of users on LGBTQ+ dating apps like Grindr (Li et al., 2017). Despite dating apps slowly taking steps to mitigate harm, such as OkCupid making users take anti-harassment pledge stating that they wouldn't send unsolicited nude photos on the app, Bumble campaigning in US and UK against such cyber flashing and getting laws against cyber flashing passed in various states in the US (Thompson, 2022), as HCI researchers, more works needs to be done in developing and implementing computer mediated solutions to prevent harm.

Risk of harm due to the institution/companies that own dating apps predominantly involves privacy breaches. Users are often unaware of how much data is being stored, shared and accessed during app use and apps have been known to violate policies and continue user data tracking and selling despite user non-consent (Brandtzaeg, Pultier, & Moen, 2019, Ghorayshi & Ray, 2018). Dating apps often claim to store location and chat data for forensic analysis to look for evidence in case of dating app related crimes (Farnden, Martini, and Raymond Choo, 2015) But research into data security employed by said companies has revealed that hackers and offenders are often able to access sensitive, stored data for malicious purposes because of improper security (Farnden, Martini, and Raymond Choo, 2015). Methods discussed that can be involved in accessing sensitive data breaches due to improper data collection, transfer, storage, etc. by companies includes Man-in-the-middle attacks (Shetty et al., 2017) and triangulation, the latter of which may not even require hacking and is in fact somewhat legal (Hoang, Asano & Yoshikawa, 2017). Researchers have been asking for policy reforms that make online platforms liable for such breaches as well as improve privacy by design and transparency of personal data flows in mobile apps. (Brandtzaeg, P. B., Pultier, A., & Moen, G. M. (2019)) to help improve user understanding and knowledge on how collected data is used ((Lutz, C., & Ranzini, G. (2017)).

Related to the institution are also the design decisions behind the application interface. From consent practices not being well supported in app designs (Im et. Al., 2021)) when they are a latent motivation behind app-use (Zytko et. Al., 2021) to constant comparisons through pictures and personal information abetting self-image issues among users, and improper distribution of affordances to protect privacy, design of these apps is often lacking in creating a safer environment for seeking relationships online. Some location-based dating applications (e.g., Tinder) allow users to hide their proximity from other users but they require users to pay to do so. This can present a barrier to users, particularly college students, who may lack the financial resources to safeguard their physical location in this manner. (Shetty et al., 2017)

They are also notorious for encouraging addictive tendencies, making it difficult for users to self-regulate app-use, especially a problem with the “swiping” affordance in dating apps, such as in Tinder, where users can swipe left or right on “matches” to reject or like respectively based on profile pictures and minimal details (Coduto, Lee-Won & Baek, 2020). Such quick, shallow dating app affordance arguably supports what can be seen as masculine values through the lens of sexual strategies theory (Buss & Schmitt. 1993.)

AI and online dating safety

AI as a design implementation for safety in dating apps has only been seen so far for scam detection and chat moderation in dating apps even though AI has long been a part of its functioning for user discovery (Dash Technologies Inc, 2022).

For automating spam detection, online dating sites employ human agents who are aided by automatic programs that narrow down the number of possible scammers on the app. However, existing automation’s low degree of accuracy does not justify a completely automated system (Kath Albury et al., 2021)). In research, Kahveci (2019) for instance, presented tools for automatically identifying malicious accounts on dating apps.

Other AI implementations for harm mitigation include Private detector AI on Bumble and Badoo which attempts to protect users from sexual harassment (“unsolicited dick pics”) by detecting and blurring nude images sent to users in the app while Tinder has attempted to use AI for chat moderation, improving through user feedback in detect problematic messages.

User Uncertainty Reduction Strategies and Impression Formation in Dating Apps

It is important to discuss users' strategies in using dating apps to better inform the design of AI for assessing risk of harm.

Users have long been attempting to engage in uncertainty reduction strategies and impression formation in online dating apps. However, they are often unsuccessful in assessing risk of harm despite employing these strategies (Centelles, Powers, & Moule, 2021).

Research has been interested in capturing self-reported risk mitigation and uncertainty reduction (Gibbs, J. L., Ellison, N. B., & Lai, C. H. (2011), Lauckner, 2018, Newlands, Newlands, et. al, 2022) strategies to understand how users go about trying to prevent harm that is intentionally or unintentionally facilitated by dating apps. Gibbs, J. L., Ellison, N. B., & Lai, C. H. (2011) in their study, generated a conceptual model of uncertainty reduction strategies and self-disclosure and outlined how users acquired the data required for their risk assessment in online dating (asking the user questions, saving online conversations and checking for consistency in narrative, verifying personal information using search engines, etc.).

Users on gay dating apps employ gay status visibility management strategies that vary in severity according to the cultural context of their country. For example, users of the gay dating app Grindr in India are forced to use extreme measures because of a close quarters living situation that range from introducing MSM partners to family (including heterosexual spouses) as friends to prevent accidental revelation/outing later (Birnholtz et. Al., 2020), hiding the application from accidentally being viewed in their smartphones (lock phone, not use in presence of friends and family, uninstall app from phone repeatedly as sharing phones with others is a common practice in India) (Birnholtz et. Al., 2020, Miao & Chan, 2021), and being careful about proximity of users and identifiable personal information used on profiles and interactions (using grandfather’s phone to register, not meeting anyone in a certain radius from self, etc.) which is also consistent with users of the Chinese gay app Blued.

HIV and STI status disclosure management is another important point of focus for gay users. Self-disclosure of HIV status in dating profiles and Sero-sorting, which is only having sexual encounters with people who have the same HIV status as the user, is an increasingly common sexual practice that can help reduce overall incidence of HIV (Cassels, S., 2009), latter of which also helps avoid fetishizers.

Transgender and disabled users use similar self-disclosure strategies, preferring direct, proactive disclosure of trans and disability status (especially obviously visible disabilities) motivated by desire of safety, certainty, (Julia R. Fernandez and Jeremy Birnholtz. 2019) and to assess people’s reactions, where a person’s positive reaction to their status is a prerequisite to pursuing a relationship and limits the potential harm of disclosure to online modalities (Julia R. Fernandez and Jeremy Birnholtz. 2019). On the other hand, non-disabled users have different rationales behind wanting to know a person’s disability status beforehand. From needing to make accommodations, so better to know early on to facilitate that, to considering one is “hiding” behind the platform when choosing not to disclose (Porter et. Al. 2017) making it difficult to affirm trust in the stranger. For better supporting self-disclosure (Handel, J., Shklovski, 2012) suggests providing options for managing levels of self-disclosure in a social network profile such as a “ASK ME” affordance for negotiable and/or sensitive information.

Strategies to detect catfishing and deception include being wary of profiles with old-looking photos that are grainy or photos that don’t show the person’s face, preferring user profiles linked to other social media, asking potential partners to video chat before meeting to confirm their identity (Lauckner, 2019). In general, because users are so wary of being deceived, they tend to make over attribution based on minimal clues provided by other online daters, scrutinizing even the most minuscule of details in order to make an assessment about viability (Ellison, 2006) but perhaps that leaves room for self-doubt and manipulation, leading to failed risk assessment. Self-disclosure about wanting to “avoid drama” and seek harmonious interactions conveys seriousness to make things work and establish ethos. (Manning, 2014).

Method

We conducted an IRB-approved participatory design interview study with 20 woman-identifying stakeholders in the United States and Canada to discuss their existing strategies for reducing uncertainty about risk of harm in online dating and then translate these strategies into mental models of inputs, decision rules, and outputs that could serve as the basis for an eventual risk awareness AI model. Participatory design is a design principle or philosophy that addresses the need to involve users who are experts in their own practices and allows us to empathize with and meet their design needs (Zhang & Wakkary, 2014).

Participant Recruitment

A total of 20 woman-identifying individuals participated in the study, who were recruited through word-of-mouth campaigns amongst the social circles of the woman-identifying members of the research team and online advertisements through social media and email lists associated with our university. Researchers employed individual methods of contacting people in their social circle that fit the Participatory requirements of the study-identifying as a woman and having some experience in online dating. They were incentivized with the potential to enrich social matching apps experiences and gift cards for participation (20) as well as recruiting other participants (4).

Ages have ranged from 18 to 41. Participants identified as black (6), white (6), Asian (4), mixed race (2), and middle eastern (1). Participants resided in 11 different states around the United States, and one resided in Canada. Interviews ranged from 1 hour and 51 minutes to 2 hours and 19 minutes.

All participants had some experience with online dating, though not necessarily through a dating application. 17 participants had experience with online dating using social matching or dating apps as well as social media platforms while 3 didn’t, using apps such as Instagram app, Patio app, and Snap Chat app for online dating instead. Tinder was reportedly the most popular (n=9) followed by bumble (4), Instagram (4), OkCupid (3), Hinge (3), Meetup (2), Facebook dating (1), Facebook (1), Snap chat (1), patio (1), CoffeeMeetsBagel (1), Muzmatch (1), match.com (1), Christian Mingle (1), eHarmony (1), and Zoosk (1).

Precautions for Participant Comfort

Precautions were taken to put participants at ease during the interview by ensuring they were aware they could stop the interview at any time and refuse to participate further without financial or social implications. The presence of male researchers during interviews was disclosed clearly early on with the affordance to ask them to leave at any point if the participant felt uncomfortable in their presence. Most of the interview was conducted by the women-presenting researchers. Participants were assured that the confidentiality of their personal details would be maintained to the highest degree possible to researchers but also informed that there was always some risk of data leaks so they could make a fully informed decision when consenting to the study. To help participants to feel safer sharing sensitive data and experiences, they were encouraged to use aliases during the interviews.

Data Collection and Analysis

Interviews begin by having participants broadly reflect on their experiences with using dating apps, followed by encouragement to expand on a particular experience that they found risky or that culminated in harm. This is used as a springboard for the participant to articulate what qualifies as harm, followed by the strategies that they practice—if any—for assessing such risks. For this purpose, we first walked participants through a storyboard scenario () to assist understanding of the concept of social matching applications followed by dummy profiles of social matching applications to facilitate understanding of opportunity or event based social matching where people are matched together through event opportunities. This is because dating apps have emerged as platforms supporting goals that go beyond dating (Birnholtz et. Al., 2015), prompting the onset of social matching apps, which are systems that recommend people to people and foster a variety of non-dating goals (Loren Terveen & David W McDonald, 2005). Modern use of dating apps is driven by far more diverse social motivations, indicating that they are already largely perceived as multi-purpose social matching technologies (Timmermans & Courtois, 2018).

They are then introduced to mock opportunity scenarios prompting participants to elaborate on how they think the system may have reached a particular conclusion about the degree of risk exhibited by that mock opportunity, thus eliciting their own personal AI models of risk.

Figure 1. An example of a storyboard scenario to help participants understand what a social matching application does.

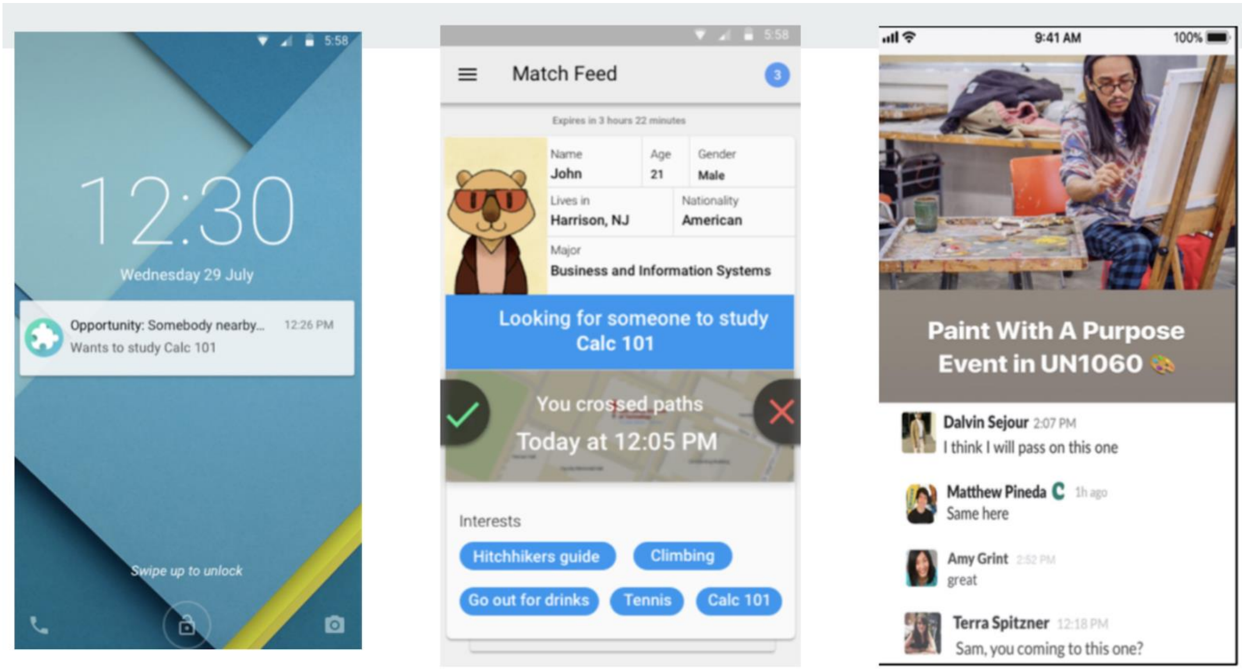

Figure 2. Dummy event profiles and matches to facilitate understanding of event opportunity based social matching apps.

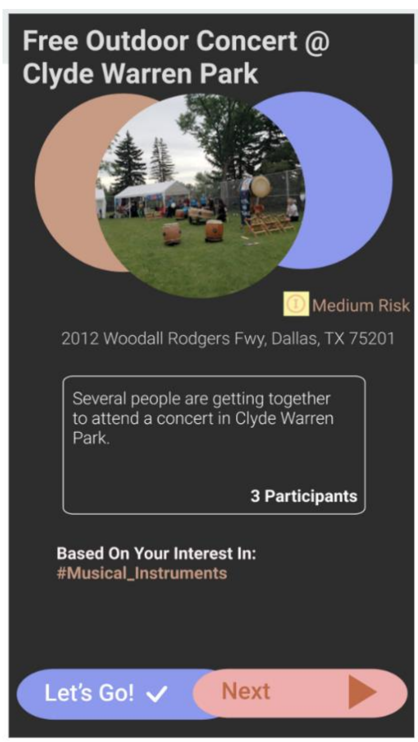

Fig 3. A mock opportunity presented to the participants.

While the participant unpacks their conceptualizations of risk and strategies for risk awareness the research assistants visually document corresponding information into what we call a “human intelligence model” (in contrast to artificial intelligence models). This carries inspiration from Lee and colleagues’ approaches to participatory AI model building (Lee et. Al., 2019., Lee et. Al., 2021). We recurrently ask the participant for feedback on the model.

All interviews were audio recorded and transcribed. The transcripts were then subjected to open coding in which we apply line-by-line coding in Dedoose and organize quotes via Miro Board to determine emergent themes around women’s strategies of reducing uncertainty of harm. This part of the analysis is exclusively qualitative in nature.

We unpack the 20 personal risk assessment models into a heat map using excel sheets to assess the features suggested by participants for assessing risk of harm and understand the way they were weighed to determine decision rules. We then use interview transcripts to elaborate on, explain, or validate at times the information captured in these personal models and come up with the three themes of rationale behind the features which explain what the participants were trying to reduce uncertainty about with certain features. The analysis of these models was therefore conducted through a hybrid quantitative and qualitative approach.

Findings

The findings first discuss the conceptual role for risk assessment AI that women favored in the online dating user experience of women. These are based purely on a qualitative approach to analysis.

We then unpack the 20 individually created directly explained models for risk assessment AI and analyze the themes behind their design, that is what information risk assessment AI should collect and how it should make risk assessment according to that information. This section is the result of the hybrid qualitative and quantitative approach to analysis mentioned in the methodology section.

How Women Envision Using Risk Assessment AI (Conceptual Role for risk assessment AI)

Risk assessment AI should not replace women’s risk assessment on dating apps completely. Conceptually, AI in this context would not be responsible for accurately automating detection of harm but be employed as an information gathering agent to help women better assess the possibility of risk of harm from not just online interactions on dating apps but also in-person meetings. P11 stated what role she wanted the “app” or AI to play in the risk assessment as:

P11: “*Yeah, if it's going to have a gradient, like 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, unless it's gathering in its, I've given it information, such as what I'm comfortable with”*

Firstly, though not explicitly expressed by everyone, there was the general idea that the “application” or the AI would have access to information about users that the participants didn’t. While some displayed a desire to be presented with this information:

Participant: *“There are some things about him that the app knows that I don't. If the app can provide me with those sets of information, then that’d be really nice.”*

Others worried about the ethical implications of how the system was gaining access to personal data that wasn't being presented by the users themselves.

P14 who worried if the image of a beer bottle in a presented mock profile was being considered an indicator of alcoholism and if not then was personal data that the app shouldn’t have access to being used to decide this user was an alcoholic, leading to the high-risk marker being placed on said profile, as she couldn't find any obvious indicator of why this user may be marked high risk.

Participant: “*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 you're collecting information on people like we have deduced they're an alcoholic.”*

P14’s confusion on how the AI reached a conclusion that she couldn't really find a rationale behind makes evident the need for interpretable outputs, as is echoed by other participants, who want to see how the “system” is reaching a certain conclusion about how “risky” a user or event is.

Interviewer: “*When you see this sign, do you care like to press on the sign and to see what the information that the application collect about this opportunities to reach this assumption of medium risk, or you will take it in general without going to…*

Participant 13: *“I would not base whatever my decision is gonna be, so if I go or if I don't go, [solely] on what the app detects as a medium risk for that event.”*

Participants also appeared to worry about AI potentially misconstruing cues and over exaggerating them leading to not processing them accurately to represent their own intuitions. They seem to feel like AI lacks the intuitiveness that human assessment often involves, for example judging a person’s online profile to likely be fake because it appears too good to be true may be a difficult cue for AI to process. As P17 describes a profile she had seen before the interview, which had images of someone flaunting their wealth in their profile pictures, the impression the person garnered from the participant was that they could have a nefarious purpose for which they are trying to lure women who may be financially needy into interacting with them.

Participant: *“One of the pictures that was used [in a profile] was of a man holding a $25 million check that he had won from the lottery. And I thought, Oh, no, this is like somebody posing as someone else. So that they can have all these women who need money [...] send them their bank account information.”*

This was not translated into the feature based personal model of the participant, either as a limitation of the study or because the participant could not find a way to translate this into a measurable form.

Doubts about the inaccuracy of such a risk assessment AI seem to come from a place of not fully understanding the limitations and especially, capabilities of AI as opposed to being averse to a new technological implementation such as AI automation for safety. P14 was, for example, hesitant about sharing criminal history with the AI system and felt more comfortable with sharing it directly with the users. She worried that the AI system may consider someone prosecuted for minor misdemeanors equally as risky as someone who may be a sex offender.

Participant: *“You know, so I guess it would my concern would be like, how does an application monitor like the severity of crimes to for like a risk assessment if that makes sense? Because I feel like that can be seen as like very unfair because like misdemeanors are, yes sure crimes but like a lot of them are really not that bad.”*

To address non-understandability, Participants suggested being presented with a list of all the features and their states that led to the AI’s output for any given opportunity.

Participant 12: *“I would actually like some information. In general, I would actually like the application to provide me with the information, to provide me with reasons why they are maybe low risk, high risk, medium risk.”*

Here the output is considered a Risk Score calculated based on the weightages that participants assigned to various feature states in their personal models. Feature state, for example, refers to the Age feature being in the “18-45 years” state for a meeting, that is the users’ age in a one-on-one meeting is between 18-45 years, which P2 scored a 7 on a self-specified 1-10 (10 is most risky, 1 is least risky) scale. The total risk score, from adding all such feature state scores relevant to an opportunity, needs to hence be explained to the participants, through the application interface.

They also want to be given an option on whether the degree of risk (low, medium, high risk or as defined by the participant) that corroborates with the total risk score is displayed on the event or user profile. Again, these are all participant-defined in the study and varying. Some participants agreed to the use of “low, medium, high risk” and/or using a “Green, yellow, red” color system while others were averse to the idea of seeing the word “Risk” being used anywhere on a profile at all. It immediately ruined her impression of said profile even if the degree of risk was “Low Risk”.

Participant 7: *"Just seeing the word risk at all [on the profile], makes me not want to go [attend the event]."*

At any rate, their worries about the accuracy of such an AI system and desire to be presented with information regarding all the features that were utilized for assessing the risk of harm of an opportunity translated to a perceived preference to maintain control over the final decision-making.

How Women Would Design Risk Assessment AI

Participants all together suggested 54 features that are used in their personal AI models that they co-created with the researchers.

Features	RISKIEST STATE OF THE FEATURE	P1	P2	P3	P4	P5	P6	P7	P8	P9	P10	P11	P12	P12(Group)	P13	P14	P15	P16	P17	P18	P19	P20
Range of values		1 to 10	1 to 10	1 to 10	0 to 10	1 to 10	1 to 3	0 to 5	0 to 10	1 to 10	1 to 5	0 to 10	1 to 10	1 to 10	1 to 10	1 to 5	0-5	1 to 5	1 to 10	1 to 10	1 to 5	1 to 3
Number of participants	one-on-one	9	9	10	7	10	3		6.5	6	2		8		6	5		5	9	6	4	3
Location familiarity	never been	9	7	10			3	3	7		3	8	9	8	10	4	4	5	6	8	4	2
Time of meeting	night	10	8	7.5	9.5	8.5	3			6			8			3	5	5	9	10	5	3
Bystander presence	none	10	9	10	10	10				8		10	7			4.5		5	8	9	4	3
Reputation of location (crime rate, reviews)	crime has occured recently or happens frequently in the area		9		8	9.5				8	4.5	9.5	10	10		3		5		8	5	3
WiFi/cell service	none				10			5	9	10	5		10		10	5	5	5				3
Familiarity with participants	none	10	9		6	7.5		3			3				7		4	5				3
Gender ratio of participants	100% male		7.5		10		3		7		4				8	5	3	5	8		4	
Person has criminal record	has a violent conviction or felony charge		9.5		10	10						10	9		10	3	4					
age of the other person	Older Men (10 years or older)		7	7.5	10				10	10					6						3	
Location provides security people/police station proximity etc.	none				9			3	5.5				9	7	8.5		2	5	6			
Profie text info completeness /Authenticity	empty	10				10	3		10		4	9.5										3
Location proximity to home	"far away"	8			7.5		3		6		2.5				7.5							
Participant is alone	is alone			7.5							3	7							8			3
Alcohol involvement at the event	present				9.5		3		7.5							3	5					2
Participant's background/rep/review	majority negative reviews	10						4				9.5			8							
Publicness of location	private residence			10			3			9						5	5		7			

Figure 4.1. The heat map is a visual representation of the 20 personal AI model women co-created in the interview session and thus represents 54 unique features. The colors taper from red to lighter colors (orange to yellow to green) as the relatively highest scoring for the riskiest feature state decreases, indicating which features weigh the most in their decision making and which weigh the least.

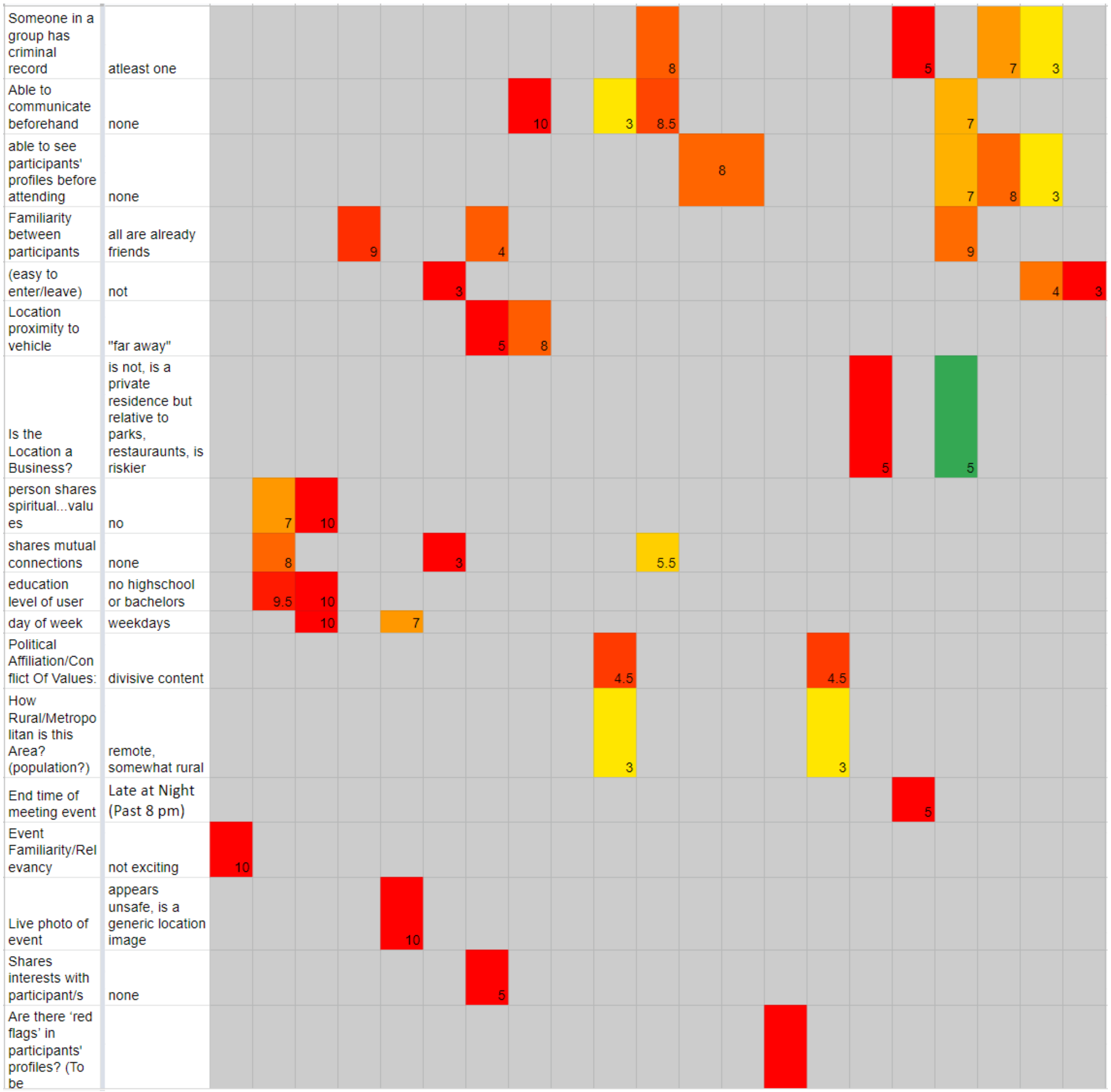

Figure 4.2. The heat map is a visual representation of the 20 personal AI model women co-created in the interview session. The colors taper from red to lighter colors (orange to yellow to green) as the relatively highest scoring for the riskiest feature state decreases, indicating which features weigh the most in their decision making and which weigh the least.

Because they have been suggested at all, each feature in these personal models is considered important for evaluating presented cues for assessing risk of harm as they all carry some weightage, however small. This translates to almost each feature associated with an event adding some risk of harm to the overall perception of risk of harm of that event. A 100% safe state of a feature is rare.

Overall, these features indicate that participants attempt to reduce uncertainty about a diverse range of harms by reducing uncertainty about three major things that could contribute to creating a safer communication environment to interact with someone for the first time in person and otherwise: *Ability to seek help, Ability to eject physically from a situation, and Gauge likelihood of users' they match with inflicting harm online or offline.* It is important to note here that participants don't only evaluate the risk of infliction of harm but also whether they are able to respond appropriately to mitigate or prevent harm in-person.

Expectations of ability to seek help immediately

Features	Possible Data Source
Number of non-participating people and families	Possibly crowd sourced or decided based on type of location (A popular park may have plenty of families in the vicinity, a cafe may have many customers and staff around)
Familiarity with participant(s)	Possibly based on mutual connections on online social platforms, users' personal assessment
Number of participants, participant is alone, is it one-on-one	Event Profile that shows number of “Checked in users”, the user
Gender ratio	Event Profile, Type of meeting (one-on-one platonic/romantic date vs group event)
Location familiarity	Self-Reported
Wifi/cell service	Crowd Sourced Reports from others who have been to the location in question
Security personnel present (bouncers, guards, etc.)	Based on the type of location (clubs, restaurants, businesses may be expected to have security), may be based on crowd sourced data (reviews, reports online)
What is the location (Publicness, is it a business, rural/metro)	Online Reviews, Online Maps

These are features that grant participants the ability to seek help immediately during a given meeting or group event. They range from the number of non-participating people or families present at the event location, how familiar they are with other user/s at an event/meeting, whether they have anyone familiar accompanying them to a group event or not to wifi/cell connectivity at the meeting location and proximity of security personnel who can intervene. Even time plays an important role as certain time ranges, such as early mornings or late nights see locations becoming barren of bystanders who may be able to assist if the participant needs help.

P3 explains her aversion to meeting someone on week/workdays: *"[On a week day] people will be busy at work. So even when I needed some help, a lot of people will be at their offices or busy or wherever working. And even my friends will not be available for me to keep checking on me. But weekend everybody's at home or at least outside in the park."*

Similarly, a user being affiliated as a student to a university for P_, and gender ratio of the users attending a group event, for others, decides, owing to some sort of gender (or student status) based camaraderie, whether or not participants can expect help from other event/meeting participants. As does positive familiarity with attending user/s or bystanders, such as regular staff at a business. Familiarity with the location of the meeting can mean that there is regular staff at the location, as mentioned before, who may be able to help.

One simple culmination of these features is questioning the "publicness" of the location of an event, which is characterized by, though not always concrete, whether it is a public location or a private residence and whether it is in a rural area or a metropolitan area like cities.

There is a trend of generalization seen where public locations and metropolitan areas are considered populated with more non-participating bystanders to turn to for assistance than a private residence or a rural setting. Of course, too many people, that is a large crowd, is easily just as unwanted as it's easy for people to intentionally or not, look past certain occurrences or for someone with disabilities or sensory issues to become uncomfortable in larger crowds.

As P16 puts it, "*I think having too large [group size], that's just as risky as having too small.*"

Ability to get to safety physically

Features	Possible Data Source
Proximity to home	Maps, user input meeting/home location
Proximity to vehicle/train	Maps, user input meeting/vehicle/public transit location
police station proximity	Online Maps (google maps, etc.)
Easy to enter/leave	Location type (A library or mall may allow more mobility than a quiet restaurant or private residence, a park is more open than a cafe, etc.)

These are features that grant or reassure participants of their ability to get to a safe place physically in case they need to. The most important feature is the location of the event being such that it grants the ability to move freely, or leave if the participant feels uncomfortable, without awkwardness or questions, and without someone being able to obstruct them in some fashion. An open, public space, such as a park, is one such a place that participants found they could easily leave.

P11: "*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."*

Other features include being at a location close to home or equivalent place of residence where there may be people who would help unquestionably, such as family members, roommates, neighbors, etc. or being able to physically barricade themselves from harm in a comfortable safe space-such as their car. One another such safe space mentioned was the police station or equivalent, or some place with security personnel who would be willing to provide assistance or protection to the participant, provide a safe shelter, or escort to the aforementioned places. Proximity to public transportation, the train station, was similarly important to P_ as being close allowed for a quicker getaway to home.

Ability to gauge user/s' likelihood of perpetuating harm by getting a sense of their personality and character

Features	Possible Data Source
Attendance rate and number of events attended	User event history captured in app, feedback from event hosts, check-in feature for users logged in user history
Education level, Cultural background, Religion, Political views, Gender, Marital status, Student affiliation	Profile Biography and/or chat conversations
Profile text completeness	Profile Biography (check for authenticity, genuineness)
Age	Profile Biography, chat conversation, ID verification (also can be used to verify authenticity of age and person)
Criminal record	Online criminal records
Mutual acquaintances	Linked social media
Reviews from others About users	Application forums/user profile rating spaces
Alcohol use, Event reviews, Event live photo Event host	Application affordances, Event profile, uploaded to the app by event hosts/participants
Can communicate beforehand + Able to see profile before attending	Application affordances
Mental health status, Drug use	Chat conversations

Participants often employed third party reviews and data to gauge the quality of the user, that is gauge the likelihood that they may inflict harm in some form. This entails looking for particular character traits they deem risky through various resources such as mutual friends, social circles, online location reviews like google or yelp reviews to check how safe the location is which could attest to certain personality traits of users, app and user reviews from other people, which, when translated into a model feature, considers how many reviews have been left for the user and what percent of them are “positive”. Chatting with the user on the dating app to gauge the same is common and preferred. Participants suggested a range of app implementations that could automate consolidation of this information or at least make it easier to access and validate users’ existence. Examples of suggested implementations include using user IDs for verification likened to Airbnb’s verification system, Live event videos posted by participants and hosts, user check-ins at live events to validate attendance and thus credibility, and review systems that account for what percent of reviews from other people are positive and negative.

An important observation that emerges from these features is that they are dependent on each other. For example, the number of participants attending an event and participants’ familiarity with the meeting location combined with it being a great distance from a safe space, such as home, can greatly impact their feeling of safety and adversely affect their decision to attend for most of the participants. That is, the risk of an opportunity is rarely the result of a single feature trumping all.

Thus, the degree of risk of harm of any given opportunity is a culmination of the probability of risk of harm that is associated with all related features.

At the same time, it’s important to note that the frequency that a feature is suggested by participants is not at all related to how important that feature is overall, indicating the **need to retain subjectivity and variability of these models in any real-life implementations.**

To achieve automated risk scoring Participants suggested profile content scanning, chat history and linked social media scrutinization as well as identity verification by linking an official id with the application. The identity verification system is likened to what is implemented in Airbnb and is likely to make sure there is only one active account linked to an ID and that it is a real person, whose personal details have been collected by the system and can be identified for prosecution should they inflict any form of harm.

Discussion

In this study, our contributions include women’s personal mental models for risk assessment, which outlines strategies women already use in their lives, how women anticipate using AI for risk assessment and how AI for risk detection should be designed in the context of online dating apps.

Participants’ strategies for uncertainty reduction have long been the focus of HCI researchers and the themes surrounding women’s risk mitigation strategies in this study are consistent and similar with previous research into users’ conceptualizations of risk in online dating (Newlands et. Al., 2022). Therefore, we do not consider this a particularly novel contribution.

As Walther et. al. report in their study, regarding information provided by someone other than the user as more valuable than that provided by the user and expecting an individual’s social network to address inconsistencies in self-representation, are some examples of strategies women in our study also addressed (Walther, 2002). However, women’s desire to be in control of the final decision contradicts critique about Bumble’s women first approach (Urszula Pruchniewska., 2020) for putting all the responsibility of preventing victimization on the user.

The study is built on the concept of augmenting women’s own strategies for uncertainty reduction, which contrasts with research that reports that women face sexual violence despite their own strategies for uncertainty reduction, which are therefore not particularly effective in preventing harm (Centelles, Powers, & Moule, 2021). We argue that given the right support, such as implementations like ID-verification and crowdsourcing data to validate a user’s or an event’s reputation and authenticity, reinforced by AI systems that present quick interpretable suggestions that can also help users to find genuine people faster, can greatly augment their strategies for safety.

We present guidelines on how dating apps and any other social matching applications that support goals beyond dating, can be designed to better foreground women's safety by pushing for personalization-friendly AI systems in dating apps that work in a hybrid setting with the human participants.

Future work can involve constructing prototypes with women as co-designers and real-life implementations to test the effectiveness of such a system in the real world. However, when the user has little trust for the AI assistance in a Human-AI team system, they do not properly rely on it, leading to a sub-optimal performance that is worse than AI or human acting alone. Same is the case for over reliance on AI (Cao & Huang, 2022.). Therefore, care must be taken in the design method of a prototype such that participants and users are able to establish an appropriate level of trust with the system.

Also, important to note is that perspectives are largely shaped by personal experiences and each person has different ones. Therefore, a real-life implementation of AI for risk assessment strategy augmentation needs to account for user variability and subjectivity and allow users to train their own models, in a way that does not require researcher involvement.

Methodological contribution: Participatory design for AI systems

The study also contributed to the methodology for conducting a PD study for designing AI systems. As an emerging field, we find it important to consider the methodology used in this study in a way that would benefit future work in co-designing AI systems with the average user.

Transparency and "explain ability" and interpretability of AI systems are two of many different things that are important to address when designing AI (Siau & Wang, 2018). By helping end users understand a system's reasoning, intelligent agents such as AI systems for risk assessment may elicit more and better feedback from users on how they can be improved, thus more closely aligning their output with each user's intentions. (Kulesza et. Al., 2012) This can be achieved by involving end users in the design process by utilizing the participatory design principle. Though not the initial purpose in this study, co-creating AI models with participants can help build participants’ trust in an AI’s workings and can help establish an appropriate degree of trust towards AI for optimal human-AI hybrid performance that can surpass their individual performances.

Further studies can also build on the concepts and methodology influenced by the Participatory design principle in this study in other contexts for more perspective on the safety needs of other, vulnerable, dating app users, such as LGBTQ+ and disabled users or minors (who have been known to use dating apps for sexual and community exploration when isolated from the urban LGBTQ+ culture, rather than to engage in illicit relationships Miao & Chan, 2021) or in the context of a different cultural setting outside the US (a different country) given that culture has an immense impact on user needs.

References

World Health Organization. (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.

MacKinnon, R. (1997). Virtual rape. Journal of Computer-Mediated Communication, 2(4), JCMC247.

Henry, N., & Powell, A. (2015). Embodied harms: Gender, shame, and technology-facilitated sexual violence. Violence against women, 21(6), 758-779.

Douglass, C. H., Wright, C. J., Davis, A. C., & Lim, M. S. (2018). Correlates of in-person and technology-facilitated sexual harassment from an online survey among young Australians. Sexual health, 15(4), 361-365.

DeKeseredy, W. S. (2020). Understanding the harms of pornography: The contributions of social scientific knowledge. Culture Reframed, 1-16.

Rubin, J. D., Blackwell, L., & Conley, T. D. (2020, April). Fragile masculinity: Men, gender, and online harassment. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-14).

Alsoubai, A., Caddle, X. V., Doherty, R., Koehler, A. T., Sanchez, E., De Choudhury, M., & Wisniewski, P. J. (2022, November). 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 (pp. 197-200).

National Crime Agency. (2016). Emerging new threat in online dating: Initial trends in internet dating-initiated serious sexual assaults.

Lindsay, M., Messing, J. T., Thaller, J., Baldwin, A., Clough, A., Bloom, T., ... & Glass, N. (2013). Survivor feedback on a safety decision aid smartphone application for college-age women in abusive relationships. Journal of Technology in Human Services, 31(4), 368-388.

Maeng, W., & Lee, J. (2022, April). Designing and Evaluating a Chatbot for Survivors of Image-Based Sexual Abuse. In CHI Conference on Human Factors in Computing Systems (pp. 1-21).

Vogels, E. A. (2020). 10 facts about Americans and online dating.

Wang, H., Cai, C., Philpot, A., Latonero, M., Hovy, E. H., & Metzler, D. (2012, June). Data integration from open internet sources to combat sex trafficking of minors. In Proceedings of the 13th Annual International Conference on Digital Government Research (pp. 246-252).

Razi, A., Kim, S., Alsoubai, A., Stringhini, G., Solorio, T., De Choudhury, M., & Wisniewski, P. J. (2021). A Human-Centered Systematic Literature Review of the Computational Approaches for Online Sexual Risk Detection. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW2), 1-38.

Powell, A., & Henry, N. (2019). Technology-facilitated sexual violence victimization: Results from an online survey of Australian adults. Journal of interpersonal violence, 34(17), 3637-3665.

Rowse, J., Bolt, C., & Gaya, S. (2020). Swipe right: the emergence of dating-app facilitated sexual assault. A descriptive retrospective audit of forensic examination caseload in an Australian metropolitan service. Forensic Science, Medicine and Pathology, 16(1), 71-77.

Shapiro, G. K., Tatar, O., Sutton, A., Fisher, W., Naz, A., Perez, S., & Rosberger, Z. (2017). Correlates of Tinder use and risky sexual behaviors in young adults. Cyberpsychology, Behavior, and Social Networking, 20(12), 727-734.

Thompson, L. (2018). “I can be your Tinder nightmare”: Harassment and misogyny in the online sexual marketplace. Feminism & Psychology, 28(1), 69-89.

Lauckner, C., Truszczynski, N., Lambert, D., Kottamasu, V., Meherally, S., Schipani-McLaughlin, A. M., ... & Hansen, N. (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), 289-306.

Lutz, C., & Ranzini, G. (2017). Where dating meets data: Investigating social and institutional privacy concerns on Tinder. Social Media+ Society, 3(1), 2056305117697735.

Brandtzaeg, P. B., Pultier, A., & Moen, G. M. (2019). Losing control to data-hungry apps: A mixed-methods approach to mobile app privacy. Social Science Computer Review, 37(4), 466-488.

Ghorayshi, A. , & Ray, D. (2018). Grindr is letting other companies see user HIV status and location data. BuzzFeed. Retrieved from https://www.buzzfeednews.com/article/azeenghorayshi/grindr-hiv-status-privacy Accessed on March 20, 2020.

Blackwell, C., Birnholtz, J., & Abbott, C. (2015). Seeing and being seen: Co-situation and impression formation using Grindr, a location-aware gay dating app. New media & society, 17(7), 1117-1136.

Callander, D., Holt, M., & Newman, C. E. (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), 3–21. doi:10.1177/1468796815581428

Miao, W., & Chan, L. S. (2021). Domesticating gay apps: An intersectional analysis of the use of blued among Chinese gay men. Journal of Computer-Mediated Communication, 26(1), 38-53. Domesticating Gay Apps: An Intersectional Analysis of the Use of Blued Among Chinese Gay Men

Li, D. H., Rawat, S., Rhoton, J., Patankar, P., Ekstrand, M. L., Rosser, B. R., & Wilkerson, J. M. (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), 324-330.

Handel, M. J., & Shklovski, I. (2012, October). Disclosure, ambiguity and risk reduction in real-time dating sites. In Proceedings of the 17th ACM international conference on Supporting group work (pp. 175-178).

Ma, X., Sun, E., & Naaman, M. (2017, February). What happens in happn: The warranting powers of location history in online dating. In Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing (pp. 41-50).

Centelles, V., Powers, R. A., & Moule Jr, R. K. (2021). An Examination of Location-Based Real-Time Dating Application Infrastructure, Profile Features, and Cybervictimization. Social Media+ Society, 7(3), 20563051211043218.

Knobloch, L. K. (2015). Uncertainty reduction theory. The international encyclopedia of interpersonal communication, 1-9.

Gibbs, J. L., Ellison, N. B., & Lai, C. H. (2011). First comes love, then comes Google: An investigation of uncertainty reduction strategies and self-disclosure in online dating. Communication Research, 38(1), 70-100. https://journals.sagepub.com/doi/abs/10.1177/0093650210377091

Dash Technologies Inc (2022, May 18). AI in Dating App: The Unexpected Love Affair between the Two. Dash Technologies Inc. https://dashtechinc.com/ai-in-dating-app-the-unexpected-love-affair-between-the-two/

Nicolas Rivero. 2021. Tinder is using AI to monitor DMS and Tame the creeps. Quartz (May 2021). https://qz.com/2011998/tinder-is-using-ai-to-monitor-dmsand-tame-the-creeps/

“Zhang, Xiao, and Ron Wakkary, 2014. "Understanding the role of designers' personal experiences in interaction design practice." Proceedings of the 2014 conference on Designing interactive systems. 2014.

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, 3340–3351.

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, 3906–3918.

Bonita C. Meyersfeld. 2012. The Council of Europe Convention on Preventing and Combating Violence Against Women and Domestic Violence. Int. Leg. Mater. 51, 1 (February 2012), 106–132. DOI:https://doi.org/10.5305/intelegamate.51.1.0106

Charlotte Watts and Cathy Zimmerman. 2002. Violence against women: global scope and magnitude. Lancet 359, 9313 (April 2002), 1232–1237. DOI:https://doi.org/10.1016/S0140-6736(02)08221-1

Smith A., Anderson M. (2015). 5 Facts about online dating. Pew Research Center. Retrieved from http://www.pewresearch.org/fact-tank/2015/04/20/5-facts-about-online-dating/

Diana Freed, Jackeline Palmer, Diana Minchala, Karen Levy, Thomas Ristenpart, and Nicola Dell. 2018. “A Stalker's Paradise”: How Intimate Partner Abusers Exploit Technology. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (CHI '18). Association for Computing Machinery, New York, NY, USA, Paper 667, 1–13. https://doi.org/10.1145/3173574.3174241

Chatterjee, R., Doerfler, P., Orgad, H., Havron, S., Palmer, J., Freed, D., ... & Ristenpart, T. (2018, May). The spyware used in intimate partner violence. In 2018 IEEE Symposium on Security and Privacy (SP) (pp. 441-458). IEEE.

Nithya Sambasivan, Amna Batool, Nova Ahmed, Tara Matthews, Kurt Thomas, Laura Sanely 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, ACM, New York, NY, USA, 1–14. DOI:https://doi.org/10.1145/3290605.3300232

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 Soc. 20, 12 (December 2018), 4689–4708. DOI:https://doi.org/10.1177/1461444818781324

Medew J. (2007). Callous teens escape jail for sex attack film. The Age. Retrieved fromhttp://www.theage.com.au/news/national/callous-teens-escape-jail-for-sex-attack-film/2007/11/05/1194117959942.html

Danielle Keats Citron and Mary Anne Franks. 2014. Criminalizing revenge porn. Wake Forest L. Rev. 49 (2014), 345.

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, 477–490.

Ellcessor. (2019). Blue-light emergency phones on campus: Media infrastructures of feeling. International Journal of Cultural Studies, 22(4), 499–518. https://doi.org/10.1177/1367877918820336

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, 59.

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, 2695–2704.

Veturekar, A., Shaikh, D., & Singh, H. (2022). Naari: A Women Safety Application. In Applications of Machine Intelligence in Engineering (pp. 155-162). CRC Press.

Roy, S., Sharma, A., & Bhattacharya, U. (2015, August). MoveFree: A ubiquitous system to provide women safety. In Proceedings of the third international symposium on women in computing and informatics (pp. 545-552).

Starks, D. L., Dillahunt, T., & Haimson, O. L. (2019, June). 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 (pp. 289-294).

Mohan, M., Sra, M., & Schmandt, C. (2017, September). Technological interventions to detect, communicate and deter sexual assault. In Proceedings of the 2017 ACM International Symposium on Wearable Computers (pp. 126-129).

Blom, J., Viswanathan, D., Spasojevic, M., Go, J., Acharya, K., & Ahonius, R. (2010, April). 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 (pp. 1841-1850).

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. In Proceedings of the Seventh International Conference on Information and Communication Technologies and Development, 1–4.

Sarosh, M. Y., Yousaf, M. A., Javed, M. M., & Shahid, S. (2016, June). Mehfoozaurat: Transforming smart phones into women safety devices against harassment. In Proceedings of the Eighth International Conference on Information and Communication Technologies and Development (pp. 1-4).

Yoo, C., & Dourish, P. (2021). Anshimi: Women's Perceptions of Safety Data and the Efficacy of a Safety Application in Seoul. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), 1-21.

Zytko, D., Furlo, N., Carlin, B., & Archer, M. (2021). Computer-mediated consent to sex: the context of Tinder. Proceedings of the ACM on Human-Computer Interaction, 5(CSCW1), 1-26.

Tariq, M. U., Razi, A., Badillo-Urquiola, K., & Wisniewski, P. (2019, July). A Review of the Gaps and Opportunities of Nudity and Skin Detection Algorithmic Research for the Purpose of Combating Adolescent Sexting Behaviors. In International Conference on Human-Computer Interaction (pp. 90-108). Springer, Cham.

Yurtseven, İ., Bagriyanik, S., & Ayvaz, S. (2021, September). A Review of Spam Detection in Social Media. In 2021 6th International Conference on Computer Science and Engineering (UBMK) (pp. 383-388). IEEE.

Al-Rousan, S., Abuhussein, A., Alsubaei, F., Collen, L., & Shiva, S. (2020, December). Ads-Guard: Detecting Scammers in Online Classified Ads. In 2020 IEEE Symposium Series on Computational Intelligence (SSCI) (pp. 1492-1498). IEEE.

Gillespie, T. (2020). Content moderation, AI, and the question of scale. Big Data & Society, 7(2), 2053951720943234.

Lorenzo-Dus, N., Kinzel, A., & Di Cristofaro, M. (2020). The communicative modus operandi of online child sexual groomers: Recurring patterns in their language use. Journal of Pragmatics, 155, 15-27.

Smith, A., & Anderson, M. (5). facts about online dating. Pew Research Center.

Stamova, I., & Draganov, M. (2020, September). Artificial Intelligence in the Digital Age. In IOP Conference Series: Materials Science and Engineering (Vol. 940, No. 1, p. 012067). IOP Publishing.

Warner, M., Gutmann, A., Sasse, M. A., & Blandford, A. (2018). Privacy unraveling around explicit HIV status disclosure fields in the online geosocial hookup app Grindr. Proceedings of the ACM on human-computer interaction, 2(CSCW), 1-22.

Warner, M., Maestre, J. F., Gibbs, J., Chung, C. F., & Blandford, A. (2019, May). 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 (pp. 1-15).

Brandtzaeg, P. B., Pultier, A., & Moen, G. M. (2019). Losing control to data-hungry apps: A mixed-methods approach to mobile app privacy. Social Science Computer Review, 37(4), 466-488.

Ghorayshi, A. , & Ray, D. (2018). Grindr is letting other companies see user HIV status and location data. BuzzFeed. Retrieved from https://www.buzzfeednews.com/article/azeenghorayshi/grindr-hiv-status-privacy Accessed on March 20, 2020.

Farnden, J., Martini, B., & Choo, K. K. R. (2015). Privacy risks in mobile dating apps. arXiv preprint arXiv:1505.02906.

R. Shetty, G. Grispos and K. -K. R. Choo, "Are You Dating Danger? An Interdisciplinary Approach to Evaluating the (In)Security of Android Dating Apps," in IEEE Transactions on Sustainable Computing, vol. 6, no. 2, pp. 197-207, 1 April-June 2021, doi: 10.1109/TSUSC.2017.2783858. Shetty et al., 2017

N. P. Hoang, Y. Asano and M. Yoshikawa, "Your neighbors are my spies: Location and other privacy concerns in GLBT-focused location-based dating applications," 2017 19th International Conference on Advanced Communication Technology (ICACT), 2017, pp. 851-860, doi: 10.23919/ICACT.2017.7890236 Your neighbors are my spies: Location and other privacy concerns in GLBT-focused location-based dating applications

Im, J., Dimond, J., Berton, M., Lee, U., Mustelier, K., Ackerman, M. S., & Gilbert, E. (2021, May). Yes: Affirmative consent as a theoretical framework for understanding and imagining social platforms. In Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems (pp. 1-18).

Coduto, K. D., Lee-Won, R. J., & Baek, Y. M. (2020). Swiping for trouble: Problematic dating application use among psychosocially distraught individuals and the paths to negative outcomes. Journal of Social and Personal Relationships, 37(1), 212–232. https://doi.org/10.1177/0265407519861153

David M. Buss and David P. Schmitt. 1993. Sexual strategies theory: An evolutionary perspective on human mating. Psychol. Rev. 100, 2 (1993), 204–232. DOI:https://doi.org/10.1037/0033-295X.100.2.204

Kath Albury, Christopher Dietzel, Tinonee Pym, Son Vivienne & Teddy Cook (2021) Not your unicorn: trans dating app users’ negotiations of personal safety and sexual health, Health Sociology Review, 30:1, 72-86, DOI: 10.1080/14461242.2020.1851610

Kahveci, O. (2019). An Approach For Detecting Online Dating Scams. An Approach For Detecting Online Dating Scams

Cassels, S., Menza, T. W., Goodreau, S. M. and Golden, M. R. HIV serosorting as a harm reduction strategy: evidence from Seattle, Washington. AIDS, 23, 18 (2009), 2497-2506

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 Human-Computer Interact. 3, CSCW (November 2019), 1–35. DOI:https://doi.org/10.1145/3359283

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 (AIES ’21), ACM New York, NY, USA, 12. DOI:https://doi.org/10.1145/3461702.3462628

Fernandez, J. R., & Birnholtz, J. (2019). " I Don't Want Them to Not Know" Investigating Decisions to Disclose Transgender Identity on Dating Platforms. Proceedings of the ACM on Human-Computer Interaction, 3(CSCW), 1-21.

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 (November 2017), 13 pages. https://doi.org/10.1145/3134722

Terveen, L., & McDonald, D. W. (2005). Social matching: A framework and research agenda. ACM transactions on computer-human interaction (TOCHI), 12(3), 401-434.

Timmermans, E., & De Caluwé, E. (2017). Development and validation of the Tinder Motives Scale (TMS). Computers in Human Behavior, 70, 341-350.

Kulesza, T., Stumpf, S., Burnett, M., & Kwan, I. (2012, May). Tell me more? The effects of mental model soundness on personalizing an intelligent agent. In Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (pp. 1-10).

Tell_Me_More_The_Effects_of_Mental_Model_Soundness_on_Personalizing_an_Intelligent_Agent

Footnotes

[1][Add footnotes, if any, on their own page following references. For APA formatting requirements, it's easy to just type your own footnote references and notes. To format a footnote reference, select the number and then, on the Home tab, in the Styles gallery, click Footnote Reference. The body of a footnote, such as this example, uses the Normal text style. *(Note: If you delete this sample footnote, don't forget to delete its in-text reference as well. That's at the end of the sample Heading 2 paragraph on the first page of body content in this template.)*]

Tables

Table 1

[Table Title]

Column Head	Column Head	Column Head	Column Head	Column Head
Row Head	123	123	123	123
Row Head	456	456	456	456
Row Head	789	789	789	789
Row Head	123	123	123	123
Row Head	456	456	456	456
Row Head	789	789	789	789

Note: [Place all tables for your paper in a tables section, following references (and, if applicable, footnotes). Start a new page for each table, include a table number and table title for each, as shown on this page. All explanatory text appears in a table note that follows the table, such as this one. Use the Table/Figure style, available on the Home tab, in the Styles gallery, to get the spacing between table and note. Tables in APA format can use single or 1.5 line spacing. Include a heading for every row and column, even if the content seems obvious. A default table style has been setup for this template that fits APA guidelines. To insert a table, on the Insert tab, click Table.]

Figures title:

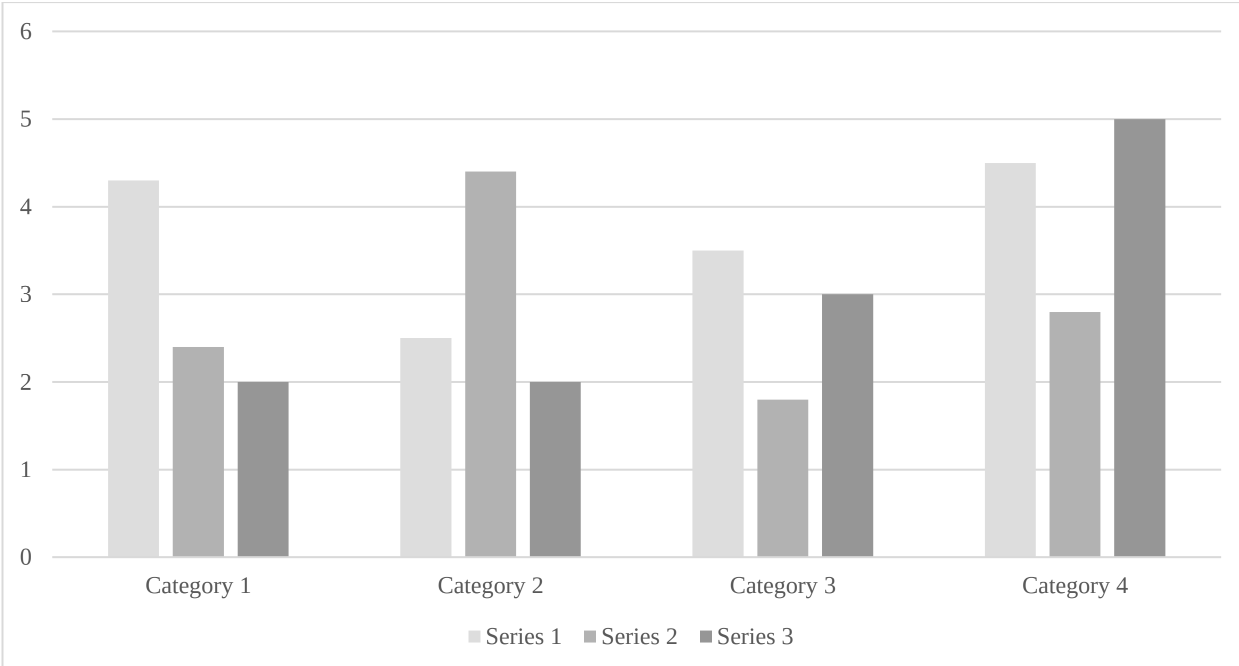

Figure 1. [Include all figures in their own section, following references (and footnotes and tables, if applicable). Include a numbered caption for each figure. Use the Table/Figure style for easy spacing between figure and caption.]

For more information about all elements of APA formatting, please consult the *APA Style Manual, 6th Edition*.