Online Dating Platform Safeguards and Self-Protection: How Dating Platforms Characterise, Respond to, and Safeguard Against Harms

Catherine O'Brien
University College London
London, United Kingdom
catherine.obrien@ucl.ac.uk

Nuur Alifah Roslan
Universiti Putra Malaysia
Selangor, Malaysia
nuuralifah@upm.edu.my

Ruba Abu-Salma
King's College London
London, United Kingdom
ruba.abu-salma@kcl.ac.uk

Steven Murdoch
University College London
London, United Kingdom
s.murdoch@ucl.ac.uk

Douglas Zytko
University of Michigan–Flint
Flint, Michigan, United States
dzytko@umich.edu

Mark Warner
University College London
London, United Kingdom
mark.warner@ucl.ac.uk

Abstract

Online dating platforms play a significant role in contemporary dating practices. While these platforms expand dating opportunities, they also expose users to harms. Through a platform-based document review, we analysed formal and informal documentation related to platform behaviours, to examine how the five most popular dating platforms in the UK characterise, address, and safeguard against harms. Our findings reveal the challenges of balancing platform accountability and user responsibility for safety, particularly as the goal of these platforms is for users to meet in-person. Platforms utilise proactive moderation tools and educational resources to enhance safety, yet many of these resources shift the burden of safety onto users. Moreover, we highlight the paradox of self-protection tools that both mitigate and enable harm, as well as identify inconsistencies in safeguarding provisions, for different geographic regions and marginalised groups.

CCS Concepts

• **Human-centered computing → Collaborative and social computing theory, concepts and paradigms**.

Keywords

Online harms; policy; online dating; terms of service; community guidelines; safeguarding, governance

ACM Reference Format:
Catherine O'Brien, Nuur Alifah Roslan, Ruba Abu-Salma, Steven Murdoch, Douglas Zytko, and Mark Warner. 2025. Online Dating Platform Safeguards and Self-Protection: How Dating Platforms Characterise, Respond to, and Safeguard Against Harms. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25), April 26-May 1, 2025, Yokohama, Japan.* ACM, New York, NY, USA, 8 pages. https://doi.org/10.1145/3706599.3719825

CHI EA '25, April 26-May 1, 2025, Yokohama, Japan

ACM ISBN 979-8-4007-1395-8/2025/04
https://doi.org/10.1145/3706599.3719825

1 Introduction

One in ten (4.9 million) online adults living in the United Kingdom (UK) visited an online dating service in 2024 [47], highlighting their role in contemporary dating practices. However, while this technology allows people to broaden the scope of dating possibilities beyond the constraints of their immediate social network [57], it also presents the risk of potentially harmful interactions. Online harms are complex, not least because of the differences in how they manifest and their severity [34, 59]. Online harms are also dynamic, impacted by sociopolitical changes, adding to this complexity. The risks of online dating platforms differ from those of other platform types such as gaming and social media which enable digital sociality as an end, whilst online dating platforms enable digital sociality to facilitate offline interactions [22]. Online dating platform harms related to sexual violence have been identified through analysis of UK crime records [1], and broader harms through interviews with online daters [18]. Surveys of users have also identified online daters' perceptions [8, 52] and experiences [21] of harms. While prior research has characterised how social media platforms define harms related to harassment [49], no prior work has looked to understand how online dating platforms characterise harms, and how they report to safeguard users against harmful behaviours. Following the UK Government's definition of online harm, the term "platform harms" is used in this paper to refer to user generated content or behaviours identified by the platforms as having the potential to cause significant physical or psychological harm to a person [27], and therefore prohibited. Following widely accepted understandings of violence more broadly [39], platform harms may be interpersonal in nature (e.g., harassment), whereas others may be self-directed (e.g., self-harm), or collective (e.g., misinformation) [75].

We report on a platform-based review of behaviour related documents (e.g., behaviour policies, safety guides), published by five of the most popular online dating platforms in the UK in 2024, to answer three research questions:

- RQ1. How do online dating platforms characterise platform harms?
- RQ2. What do online dating platforms say about how they respond to platform behaviour violations?
- RQ3. What safeguards do online dating platforms say they put in place, to keep their users safe?

Our work offers a comparative examination across dating platforms, investigating their formal and informal documents [49]. We present our analysis on the characterisation of harms, enforcement actions, awareness of off-platform impact and risks, safeguarding tools, advice, guidance, and user responsibility. We highlight the tension between platforms' responsibility to safeguard, both on and off platform (i.e., in-person), and user responsibilities for personal and collective safety, contributing to broader conversations around user empowerment and platform accountability.

2 Related Work

2.1 Harms in the Context of Online Dating

Users face a variety of potential harms on online dating platforms, which emerge through the interplay of technology [4, 14, 24, 48, 68], societal norms [38, 48, 74], and individual behaviours. These harms can impact mental health [37, 45, 67], safety [40], and overall well-being [58] of individuals. Understanding these harms requires examining the risks that arise in respect to the online-to-offline nature of interaction [22], and user vulnerabilities (e.g., LGBTQ+ individuals, racial and ethnic minorities, and individuals with disabilities [62, 73]). Unlike offline dating, which typically fosters trust through face-to-face interactions in a shared social context, online dating is characterised by computer-mediated communication that takes place at a physical distance [33]. This physical distance, and associated relative anonymity, may facilitate deceptive self-presentation [31, 50]. Scammers often exploit emotional vulnerabilities to establish trust and manipulate victims into financial transactions. Romance scams, for example, involve creating fake emotional bonds to extract money from victims, often leading to significant financial and emotional distress [72]. Moreover, the ease of access to personal information on dating apps can also facilitate stalking and harassment, posing significant safety risks to users [24]. Furthermore, sexual harm is a serious concern within online dating environments [22], especially for women, LGBTQ+ individuals, and other vulnerable populations. Users can experience harassment, unsolicited explicit content, and even physical assaults, which are often normalised within online dating culture [14]. Technology-facilitated sexual violence poses additional risks, with survivors facing severe psychological repercussions [23]. There are also risks related to hate speech and discrimination on online dating platforms. LGBTQ+ users and racial or ethnic minorities may be subjected to bias, harassment, and abuse, further complicating their dating experiences [36]. For some users, these experiences could lead to mental health challenges such as PTSD and anxiety [45]. Moreover, it is important to recognise the intersectionality of identity in online dating spaces [42], as race or ethnicity [56] and disability [74] can heighten the risk of harm, as marginalised individuals often encounter compound vulnerabilities in online spaces [16].

2.2 Online Dating Platform Safety

Dating apps have been criticized for inadequate safety measures, such as insufficient privacy controls [24, 69] and a lack of abuse prevention features, rendering users vulnerable to various forms of dating violence [40]. Features such as location tracking or the necessity of sharing personal information to make connections with others [44] can also facilitate physical harassment or dating violence [2, 5, 6, 17, 20, 26, 30, 32, 40, 41, 43, 53, 55]. However, other work problematises the notion that dating app use is inherently riskier than other forms of dating [3, 64].

Platforms have implemented automated content moderation systems to detect potentially violating and harmful content. Badoo and Bumble blur explicit images to protect users from sexual harassment, Tinder moderates chats for problematic messages, and others use automated tools to assist human moderators with detecting scams [19]. Datey and Zytko highlight the potential of AI-based risk detection technologies to assist women in assessing potential harms and risks associated with online dating, combining proactive automated detection with users' personal preferences to mitigate risks. While these approaches are proactive and easily scaleable, there are concerns related to bias [10, 12, 25] and privacy [71]. Comparatively, user reporting as a detection measure does not face the same scrutiny, with one participant in the interaction consenting to the disclosure of the reported material. Other reactive features, such as blocking and panic buttons for in-person interactions, also shift responsibility to users post-harm [5]. That being said, more proactive measures, such as concealing location data [15] or Bumble's women-first messaging rule [54], have also faced criticism for their limited effectiveness [5]. In addition to safety tools, educational initiatives have proven essential in reducing the risks of online dating abuse [66].

3 Methodology

3.1 Platform and Document Selection

Platforms were selected (Appendix Table 2) based on popularity as reported in a 2024 UK survey commissioned by Ofcom [47].[1] To select documents, we followed the method of [49], selecting formal documents that are legally binding (e.g., Terms of Service), and informal documents that are non-legally binding (e.g., safety guides). Documents across the selected platforms Badoo (n = 22), Bumble (n = 35), Grindr (n = 19), Hinge (n = 19), and Tinder (n = 23), were identified through manual website searches on each platform's online website (Appendix Table 3), and through keyword-based web searches (Appendix Table 4) of these websites. The manual search was conducted by the first author who reviewed the Terms of Service of each platform to identify references to additional documentation (e.g., community guidelines). Following this, a search was conducted using the search terms selected based on a review of prior work [49, 61, 63] and an Ofcom's experience survey [46].

3.2 Qualitative Coding Methodology

We used an inductive coding approach to collaboratively develop a codebook. Our analysis was part of a larger project evaluating platform documents across four categories (social, dating, gaming, and messaging), and so our codebook was developed using documents across all platform categories. To develop our codebook, we coded 12 documents that spanned the four categories, and included both formal and informal documents. The team met to discuss the codes to develop an initial codebook. The codebook was then evaluated

[1] Ofcom is the UK online safety regulator.

	Abuse	(Sexual) assault	Blackmail	Body shaming	Bullying	Child abuse	Derogatory	Doxxing	False reporting	(Sexual) harassment	Hate speech	Identity-based attacks	Illegal goods/services	Inauthenticity	Insult	Intimidation	Offensive	Libellous	Manipulation	Misinformation	Scam/theft	Self-harm	Stalking	Violence/extremism	Vulgarism
Badoo	FI	I	I	I	FI	FI	I	FI	I	FI	FI	I	FI	FI	FI	I	FI	FI	FI	I	FI	I	I	FI	I
Bumble	FI	I	I	I	I	FI	I	FI	FI	FI	FI	I	FI	FI	FI	I	FI	FI	FI	I	FI	I	FI	FI	I
Grindr	FI	I	I	FI	FI	FI	I	FI	I	FI	FI	FI	FI	FI	FI		FI	F	I	I	FI	I	FI	FI	F
Hinge	FI	FI			FI	FI	I	F	I	FI	FI	I	FI	FI	FI	FI	FI	FI			FI	FI	FI	FI	
Tinder	FI	FI	I		FI	FI	I	FI	FI	FI	FI	I	FI	FI		I	FI	F	I	I	FI	FI	F	FI	

Table 1: Harms recognised and prohibited within platforms, across formal (F) and informal (I) documents.

on a further 13 documents, with refinements being made collaboratively. Lastly, a deductive analysis of all dating platform documents (n = 118) was performed.[2] In this paper, we report on five of the themes developed from our analysis: (1) harm characterisation, (2) platform enforcement actions for behavioural policy violations, (3) safeguarding tools, (4) safeguarding advice and guidance, and (5) user responsibility for safety.

4 Policy Analysis Findings

4.1 Characterisation of Platform Harms

All platforms characterise harms in terms of a broad set of behaviours including abuse, doxxing, manipulation, and violence (Table 1). While some of these harms are explicitly defined by the platforms, most are not. Similarly to Pater et al.'s [49] social media policy review, no platforms explicitly define harassment. However, Bumble and Badoo do explicitly define *sexual* harassment. Bumble defines it as "any unwanted or unwelcome sexual behaviors between members." Badoo's definition is similar; it conditions that the behaviour be "non-physical." Badoo differentiates between sexual harassment as a non-physical activity, and sexual assault as a physical activity, defining it as: "unwanted physical contact or attempted physical contact that is sexual in nature." While the majority of harms are characterised as acts, some are characterised as incitement ("inciting violence against a person or group" [Hinge, safety guide]), promotions ("promotes harmful stereotypes isn't acceptable" [Badoo, reporting guide]), and threats ("sending threats or offensive messages to someone on and off the app" [Tinder, Terms of Use]). Moreover, platforms differentiate some behaviours based on their repetitive nature ("harassing you by messaging too much" [Bumble, reporting guide]) and intentionality ("intentionally misleading, false, or otherwise inappropriate" [Grindr, terms of service]). Finally, all platforms differentiate certain harms based on the target, including children, individuals, groups, those with protected characteristics, and the self.

4.2 Platform Enforcement Actions for Behavioural Policy Violations

Enforcement actions for violations of community guidelines are reported by all platforms. All platforms report account suspension and removal, content removal, and sending warnings or nudges. Tinder is the only platform to mention restricting visibility of content, and Grindr is the only platform to mention actions related to requesting users to remove their own content. Badoo, Bumble, and Hinge all mention action related to suspending off-platform accounts where possible (e.g., accounts on platforms owned by the same parent company). Finally, all state law enforcement reporting as a potential enforcement action. All platforms mention aspects of harm severity in relation to enforcement action. For example, Grindr's community guidelines state "in the case of severe violations or repeated offenses, we take more serious measures, such as permanent bans." Bumble also mentions aspects of severity in relation to identity-based hate, also highlighting considerations related to "context, intention, and impact."

4.3 Safeguarding Tools

Encouraging users to remain on the platform. As platforms employ safety tools and other safeguarding measures within-app to protect users, all except Badoo caution users against taking their interactions off the platform too quickly: "Keep conversations on the Hinge platform while you're getting to know someone [..] users with bad intentions often try to move the conversation to text, messaging apps, email, or phone right away" [Hinge, safety guide].

Criminal background checks. We found that the Terms of Service of Badoo, Tinder, and Hinge prohibit anyone convicted of violent or sexual crimes from becoming a user, and Bumble has the right to ban

[2]One document was produced by Match Group and was relevant to both Tinder and Hinge.

any user found to have faced such convictions. The Terms of Service for all platforms, including Grindr's, state the platforms' right to conduct criminal background checks. However, Grindr, Hinge, and Tinder clarify that this is not currently performed. Moreover, Bumble seeks to clarify that while they may perform some criminal background checks, this measure is not "foolproof," as they can be circumvented and rely on databases that are often not up to date.

User blocking. All platforms have developed safety tools or features that are either automatically implemented by the platform or could be applied by the user to protect themselves. On Badoo, Bumble, Hinge, and Tinder, when reporting a user, the reported account is automatically blocked, "the moment someone is reported, we make sure you never see each other's profile again" [Hinge, safety guide]. However, Grindr's reporting function requires that users separately block other users: "Keep in mind that reporting a profile does not block it" [Grindr, reporting guide].

Geographical risk safeguards. Platforms recognise that LGBTQ+ users face additional risks, and both Grindr and Tinder have safety features to protect this group. On Tinder, this includes proactively warning LGBTQ+ members of risks when travelling to countries with laws that target this community, and advising them to remove their sexual orientation from their profile while in those countries. Similarly, Grindr states that "In those areas, Grindr may automatically obscure user's locations, or even have this feature turned off completely" [Grindr, safety guide]. Additionally, Grindr offers discrete app icons to hide the Grindr logo from user devices, as well as PIN code protection. These safeguarding measures are tailored to a specific user group, and help protect from specific off-platform harms.

Screenshot blocking. Grindr and Badoo employ 'Screenshot Block', which prevents users from taking screenshots of conversations on the app or saving images and recordings to safeguard against non-consensual sharing. However, Grindr warns that this does not prevent bad actors from employing creative means to capture content, such as photographing the screen, highlighting the platforms' lack of active control over off-platform behaviours. Badoo warns users that the technology is only effective on Android, and iOS users are only prompted to reconsider their actions.

Proactive content moderation. We found that some platforms embed proactive content moderation features into their platforms. Hinge and Tinder, both owned by Match, have an "Are You Sure?" feature, which detects harmful content before being sent and prompts users to reconsider their language. While this feature places responsibility on the sender, others detect harmful content received, blurring or flagging content to protect users. All platforms employ some form of this technology, either protecting against rude or offensive language or blurring potential nude or sexual images.

4.4 Safeguarding Advice and Guidance

Working with third parties and sign-posting resources. All platforms report working with third parties (e.g., governments) to help prevent or protect against different forms of harm. All platforms use third-party content, signposting support for suicidal/self-injurious behaviours and sexual/physical abuse. Some platforms signpost external resources targeted at supporting individuals with protected characteristics associated with race (Grindr, Hinge, Tinder) or LGBTQ+ status (Bumble, Hinge, Grindr, Tinder)—whether these relate to mental health, experiences of physical violence, or hate and discrimination. However, many of these resources are geographically restricted – e.g., the external resource BrightCheck, which is signposted by Tinder, is only available in the US.

Guidance around consent. Guidance around consent both on and off platforms is prevalent across platforms and largely dictates whether behaviours related to sexually-explicit content are prohibited or permitted. E.g., while it is expressly permissible to engage in sexually-explicit conversations via messaging on Badoo, Bumble, Grindr, and Tinder, both parties must consent. However, these behaviours are not permitted in public profile content, as this does not allow users to consent to view this content. Furthermore, all platforms engage in defining and educating more broadly on what constitutes consent, particularly in relation to in-person interactions. Badoo discusses prohibited acts of sexual assault, including unwanted sexual touching, kissing, sexual penetration, or attempts to do so, and 'stealthing'.[3] Other platforms, such as Grindr and Tinder, delve more specifically into how to check in with sexual partners during intimacy, reading verbal and non-verbal cues such as "nodding, pulling someone closer, or active engagement, such as mutual touching." [Tinder, safety guide]. Moreover, there is guidance on the nature of informed and enthusiastic consent, as well as the right to remove consent, and how this should inform users of appropriate behaviour with respect to sexually-explicit material and sexual interactions on and off the platform.

Safety guidance. Multiple safety guides have been produced by all five platforms, and similarly offer advice on interactions on and off platform, ranging from cautioning users against sharing their personal information with others on the app, to notifying friends or family when meeting up with others, "If it makes you feel more comfortable, you can even send your person a screenshot of your date's profile for good measure" [Bumble, safety guide]. In this way, platforms place the responsibility of caution and decision-making around personal safety and boundaries firmly onto users, without prohibiting behaviours, such as sharing personal information via messaging.

4.5 User Responsibility for Safety

Encouraging a 'common sense' and cautious approach, and self-care. Safety is established by all platforms as in part the responsibility of the user, particularly within off-platform interactions. Bumble, Grindr, Hinge, and Tinder encourage users to apply caution and consider their safety on the app. E.g., Grindr instructs users to "use a high level of caution when chatting with others", "think critically about where to have conversations", and "be aware that some software can reverse the blur feature" [Grindr, safety guide], rendering users responsible for their interactions with others on the platform. In regions where being LGBTQ+ poses a risk, Bumble, Grindr, Hinge, and Tinder also recommend developing "a common sense response" [Grindr, safety guide] and offer additional guidance.

[3] Stealthing refers to non-consensual removal of barrier contraceptive.

Similarly, Tinder, Hinge, and Bumble stress the importance of caution, care, and judgement both on and off platform: "You agree to use caution in all interactions with other users, particularly if you decide to communicate off the service or meet in person" [Hinge and Tinder, Terms of Service]. Some platforms go further, advising users to research their matches: "It's always okay to do a little homework before meeting a new person in real life. Do they have any skeletons big enough to have made the news? Are they the webmaster for a popular Buffy fan page? It never hurts to know these things early" [Bumble, safety guide]. Additionally, Tinder encourages US users to utilise external tools to perform background checks on their matches. In addition, where harm occurs, Bumble, Grindr, and Tinder advocate for "self care" practices, urging users to consider how to respond or decide which would best help them "heal" [Grindr, community guidelines].

Encouraging safe user reporting and bystander intervention. While enforcement is the responsibility of the platform, all five platforms make clear how users are responsible for reporting other users and deciding the support that they require. Moreover, platforms place responsibility on users to decide whether to call out bad behaviour in addition to reporting it. This is expressed in Grindr's community guidelines: "If you feel safe doing so, stand up for others." A frequently mentioned method of intervening as a bystander is through user reporting. Although all platforms deploy automated detection tools combined with human moderation, user reporting is the most frequently mentioned form of detecting policy violations across all platforms. To support users in reporting and standing up for others through reporting, all platforms make reference to the "confidential" or "anonymous" nature of reporting. Moreover, all five platforms make reference to the idea that reporting holds the dual function of protecting the self and other users.

Avoiding false reporting and discrimination. All platforms recommend users to un-match or block instead of reporting where behaviour does not violate a policy. They designate false reporting as a prohibited behaviour: "If someone is making you feel uncomfortable on the app, we encourage you to block them. If they are violating our Community Guidelines, please report them" [Grindr, reporting guide]. In particular, all platforms are aware that individuals with protected characteristics, particularly surrounding gender identity, are at risk of victimisation through false reporting: "If we believe you are intentionally reporting someone because of their gender expression or another protected attribute, we may take action against your account" [Badoo, hate speech guidelines]. Therefore, while reporting is encouraged to protect the self and community, it must follow community guidelines as to what constitutes a policy violation, and it is not the user's responsibility to decide what behaviour is violating.

5 Discussion

5.1 Balancing Platform and User Responsibility

Online dating platforms present unique risks to users [40] due to the goals of these platforms being to move user interactions offline; these risks are highlighted by the platforms and present a challenge for them to balance messaging around platform and user responsibility. All platforms apply top-down safety mechanisms, such as proactive content moderation tools [60, 70, 71] to enhance safety within-app, whilst also encouraging user responsibility to help users stay safe once they move beyond the platforms' protective mechanisms. By their nature, the risks associated with meeting up with strangers following connecting on these platforms push the responsibility for safety off-platform onto users. While Lee's study on Tinder and Bumble [40] has reported limited guidance from platforms on safe dating strategies, our 2024 study reports on this type of guidance, which is common across four of the five platforms we analysed. Of particular note was the platforms' emphasis on consent in relation to personal safety. While "consent apps" have been developed as a means of mediating consent, they have also been criticised for subverting sexual agency [43, 51, 76]. By placing emphasis on education by the platforms we analysed, rather than the implementation of consent tools, platforms allow users to retain agency in their sexual encounters while promoting personal and community safety.

5.2 Tools and Techniques to Both Harm and Prevent Harm

Tensions emerge between platform guidance and safeguarding tools, as recommendations for self-protection can also be exploited to engage in prohibited behaviours that harm other users. E.g., Bumble and Grindr suggest "screenshotting" and sharing a match's profile before meeting, but doxxing or non-consensual sharing is prohibited by all platforms (see Table 1). Thus, while Grindr specifies that "you may not share other people's information for any reason (even if you mean well)" [Grindr, community guidelines], and, like Badoo, employs screenshot blocking technology, it also encourages it as a self-protective tactic. Furthermore, all platforms prohibit false reporting on the basis of identity or protected characteristics, yet Badoo, Bumble, Hinge, and Tinder use open and subjective terminology such as "uncomfortable or unsafe" as reasons for reporting, which complicates definitions of harassment and can be used to justify false reporting, whether or not there is malintent. Relatedly, language in informal documentation illustrates the platforms' struggle to balance highlighting the enjoyable user experience the platforms offer and ensuring users are informed of the risks inherent to the goals of the platforms, i.e., meeting strangers. E.g., Bumble's safety guide humorously downplays risks while advising users to research matches. This illustrates the challenge of "empowering" (Bumble, Hinge, Grindr, Tinder) users to protect their safety off-platform while not scaring them off altogether. Furthermore, there is a fine line between researching a match for safety (Bumble, Grindr) and crossing into intrusive or "creepy" behaviour (Tinder), with platforms like Badoo explicitly labelling "stalking profiles on social media" as reportable harassment. Tinder recommends background checks using external tools, but (unlike Bumble) it fails to warn users of their potential inaccuracies, creating a false sense of security that contradicts the platform's emphasis on user caution. The tensions identified between formal and informal platform policies—such as contradictions between recommended safety practices and prohibitions against certain behaviours—underscore the broader challenge of ensuring platform accountability while empowering users to protect themselves. These insights call for further dialogue on how online dating platforms can refine their policies to

mitigate against unintended harms, particularly regarding false reporting and the exploitation of safety tools. Future research should explore how platform governance can be improved to provide more consistent and effective protections for different marginalised users, ensuring that safety measures do not inadvertently reinforce inequalities in online dating experiences.

5.3 Geographical and Protective Characteristic Inconsistencies in Safeguarding

There are geographic differences in user access to external safety tools and resources, with many catering only to the US or Canada. Additionally, tailored protections vary by group. While all platforms reference individuals with "protected" attributes, only LGBTQ+ users receive tailored safeguarding tools, although groups with other protected characteristics, such as race, do receive tailored signposted external resources. E.g., Badoo, Bumble, Hinge, and Tinder single out gender identity and being trans as protected attributes that are at particular risk of false reporting. In addition, Tinder, Hinge, and Grindr address the increased in-person risks LGBTQ+ users face through their geographical risk safeguards. This is particularly evident for Grindr, which features the most extensive resources, likely reflecting the needs of their user base. Increased safeguards for this group may reflect an awareness of the vulnerable status of LGBTQ+ users on dating apps due to their increased risk of experiencing sexual violence and harassment [9, 13, 77]. In contrast, other marginalised groups, such as users with disabilities [73], are mentioned far less frequently. This raises the question of equity in platform safety design and highlights the need for more inclusive and comprehensive safeguarding strategies that account for intersectional vulnerabilities. Future work may involve collaborating with marginalised user groups to uncover their specific safety concerns, examining where these diverge or overlap, for platforms to develop tailored safeguarding measures – either through tools or guidance. For example, this may involve consulting different marginalised groups on existing self-protecting practices they engage in, both on and off platforms, which could be developed into guidance specific to the existing capabilities of platforms.

6 Conclusion

In this paper, we emphasise the complexities of balancing platform and user responsibilities in a context in which users are required to move off platform to facilitate meaningful connections. Moreover, we highlight the variety of proactive content moderation tools employed on platforms to safeguard users, and their development of educational resources that place safety responsibility on users in their off-platform engagements. This balance is further complicated by the dual affordances of safety tools – both safeguarding from and potentially enabling harm. In general, these findings emphasise the tension between empowering users to navigate risks when they move beyond the platform and ensuring platform accountability for the interactions they facilitate.

Acknowledgments

Thanks go to Andreas Gutmann (Ofcom, UCL) for his valuable comments and suggestions. This research was supported by UK Research and Innovation (UKRI) through REPHRAIN (EP/V011189/1), the UK's Research Centre on Privacy, Harm Reduction and Adversarial Influence Online.

References

[1] National Crime Agency. 2022. Sexual Offences Initiated Via Online Dating Submitted to SCAS 2003 to 2021. https://www.nationalcrimeagency.gov.uk/who-we-are/publications/583-online-dating-scas-statistics-2021/file
[2] Esma Aïmeur, Nicolás Díaz Ferreyra, and Hicham Hage. 2019. Manipulation and Malicious Personalization: Exploring the Self-Disclosure Biases Exploited by Deceptive Attackers on Social Media. *Frontiers in Artificial Intelligence* 2 (2019), 1–27.
[3] Kath Albury, Anthony McCosker, Tinonee Pym, and Paul Byron. 2020. Dating Apps as Public Health 'Problems': Cautionary Tales and Vernacular Pedagogies in News Media. *Health Sociology Review* 29, 3 (2020), 232–248.
[4] Fatemeh Alizadeh, Dennis Lawo, Gunnar Stevens, Douglas Zytko, and Motahhare Eslami. 2024. When the "Matchmaker" Does Not Have Your Interest at Heart: Perceived Algorithmic Harms, Folk Theories, and Users' Counter-Strategies on Tinder. *Proceedings of the ACM on Human-Computer Interaction* 8 (2024), 1–29. https://doi.org/10.1145/3689710
[5] Hanan Khalid Aljasim and Douglas Zytko. 2023. Foregrounding Women's Safety in Mobile Social Matching and Dating Apps: A Participatory Design Study. *Proceedings of the ACM on Human-Computer Interaction* 7, GROUP (2023), 1–25.
[6] Sima Amirkhani, Fatemeh Alizadeh, Dave Randall, and Gunnar Stevens. 2024. Beyond Dollars: Unveiling the Deeper Layers of Online Romance Scams Introducing "Body Scam". In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 57, 6 pages. https://doi.org/10.1145/3613905.3651004
[7] Badoo. 2024. *About Badoo.* https://badoo.com/en/
[8] Alyssa M Beauchamp, Hannah R Cotton, Allison T LeClere, Emily K Reynolds, Sean J Riordan, and Kathleen E Sullivan. 2017. Super Likes and Right Swipes: How Undergraduate Women Experience Dating Apps. *Journal of the Student Personnel Association at Indiana University* (2017), 1–16.
[9] Rena Bivens and Anna Shah Hoque. 2018. Programming Sex, Gender, and Sexuality: Infrastructural Failures in the "Feminist" Dating App Bumble. *Canadian Journal of Communication* 43, 3 (2018), 441–459.
[10] Tolga Bolukbasi, Kai-Wei Chang, James Y Zou, Venkatesh Saligrama, and Adam T Kalai. 2016. Man is to Computer Programmer as Woman is to Homemaker? Debiasing Word Embeddings. *Advances in Neural Information Processing Systems* 29 (2016), 1–9.
[11] Bumble. 2024. *What is Bumble?* https://bumble.com/en/help/what-is--bumble
[12] Joy Buolamwini and Timnit Gebru. 2018. Gender Shades: Intersectional Accuracy Disparities in Commercial Gender Classification. In *Conference on Fairness, Accountability and Transparency*. PMLR, 77–91.
[13] Paul Byron, Kath Albury, and Tinonee Pym. 2021. Hooking Up With Friends: LGBTQ+ Young People, Dating Apps, Friendship and Safety. *Media, Culture & Society* 43, 3 (2021), 497–514.
[14] Elena Cama. 2021. Understanding Experiences of Sexual Harms Facilitated through Dating and Hook Up Apps among Women and Girls . In *The Emerald International Handbook of Technology-Facilitated Violence and Abuse*. Emerald Publishing Limited.
[15] Vanessa Centelles, Ráchael A Powers, and Richard K Moule Jr. 2021. An Examination of Location-Based Real-Time Dating Application Infrastructure, Profile Features, and Cybervictimization. *Social Media+ Society* 7, 3 (2021), 1–11.
[16] Christopher T. Conner. 2022. How Sexual Racism and Other Discriminatory Behaviors are Rationalized in Online Dating Apps. *Deviant Behavior* 44 (2022). https://doi.org/10.1080/01639625.2021.2019566
[17] Elena Francesca Corriero and Stephanie Tom Tong. 2016. Managing Uncertainty in Mobile Dating Applications: Goals, Concerns of Use, and Information Seeking in Grindr. *Mobile Media & Communication* 4, 1 (2016), 121–141.
[18] Danielle Couch, Pranee Liamputtong, and Marian Pitts. 2012. What Are the Real and Perceived Risks and Dangers of Online Dating? Perspectives From Online Daters: Health Risks in the Media. *Health, Risk & Society* 14, 7-8 (2012), 697–714.
[19] Isha Datey and Douglas Zytko. 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. *Proceedings of the ACM on Human-Computer Interaction* 8, CSCW2 (2024), 1–41.
[20] Karel Dhondt, Victor Le Pochat, Yana Dimova, Wouter Joosen, and Stijn Volckaert. 2024. Swipe Left for Identity Theft: An Analysis of User Data Privacy Risks on Location-Based Dating Apps. In *33rd USENIX Security Symposium (USENIX Security 24)*. 5053–5070.

[21] Ashley K Fansher and Sara Eckinger. 2021. Tinder Tales: An Exploratory Study of Online Dating Users and Their Most Interesting Stories. *Deviant Behavior* 42, 9 (2021), 1194–1208.

[22] Eric Filice, Kavishka D Abeywickrama, Diana C Parry, and Corey W Johnson. 2022. Sexual Violence and Abuse in Online Dating: A Scoping Review. *Aggression and Violent Behavior* 67 (2022), 1–18.

[23] Eric Filice, Amy Matharu, Diana C. Parry, and Corey W. Johnson. 2024. A Thousand Catcalls: Survivors' Experiences of Sexual Violence in Online Dating. *Leisure Sciences* (2024). https://doi.org/10.1080/01490400.2024.2330946

[24] Rachel Fletcher, Calli Tzani, and Maria Ioannou. 2024. The Dark Side of Artificial Intelligence–Risks Arising in Dating Applications. *Assessment and Development Matters* 16, 1 (2024), 17–23.

[25] Nikhil Garg, Londa Schiebinger, Dan Jurafsky, and James Zou. 2018. Word Embeddings Quantify 100 Years of Gender and Ethnic Stereotypes. *Proceedings of the National Academy of Sciences* 115, 16 (2018), E3635–E3644.

[26] 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.

[27] UK Government. 2024. Understanding and Reporting Online Harms on Your Online Platform. https://www.gov.uk/guidance/understanding-and-reporting-online-harms-on-your-online-platform Accessed: 24 Feb. 2025.

[28] Grindr. 2024. *About Grindr.* https://www.grindr.com/about

[29] Grindr. 2024. *How Does Grindr Work?* https://www.grindr.com/blog/how-does-grindr-work

[30] Ralph Gross and Alessandro Acquisti. 2005. Information Revelation and Privacy in Online Social Networks. In *Proceedings of the 2005 ACM workshop on Privacy in the electronic society*. 71–80.

[31] Rosanna E Guadagno, Bradley M Okdie, and Sara A Kruse. 2012. Dating Deception: Gender, Online Dating, and Exaggerated Self-Presentation. *Computers in Human Behavior* 28, 2 (2012), 642–647.

[32] Seda Gürses and Claudia Diaz. 2013. Two Tales of Privacy in Online Social Networks. *IEEE Security & Privacy* 11, 3 (2013), 29–37.

[33] Lara Hallam, Michel Walrave, and Charlotte. 2018. Information Disclosure, Trust and Health Risks in Online Dating. In *Sexting*. Palgrave Macmillan, Cham.

[34] Andrés Domínguez Hernández, Kopo M. Ramokapane, Partha Das Chowdhury, Ola Michalec, Emily R. Godwin, Alicia Cork, and Awais Rashid. 2023. Co-creating a Transdisciplinary Map of Technology-mediated Harms, Risks and Vulnerabilities: Challenges, Ambivalences and Opportunities. *Proceedings of the ACM on Human-Computer Interaction* 7 (2023). https://doi.org/10.1145/3610179

[35] Hinge. 2024. *Hinge's Mission.* https://hinge.co/mission

[36] Mujlauwidzatul Husna. 2024. Challenging Gender Stereotypes in The Early Years: hanging the Narrative. *Journal of Gender Studies* 33 (2024). https://doi.org/10.1080/09589236.2023.2291929

[37] T. Jennings, Yen Ling Chen, Bailey M. Way, Nicholas C. Borgogna, and Shane W. Kraus. 2023. Associations Between Online Dating Platform Use and Mental and Sexual Health Among a Mixed Sexuality College Student Sample. *Computers in Human Behavior* 144 (2023). https://doi.org/10.1016/j.chb.2023.107727

[38] Evanthia Kavroulaki. 2021. "Congratulations! You Just Won the Title for 'Worse Tinder Opening Line'": Inappropriate Behaviour and Impoliteness in Online Dating. *Journal of Language Aggression and Conflict* (2021). https://doi.org/10.1075/JLAC.00063.KAV

[39] Etienne G Krug, James A Mercy, Linda L Dahlberg, and Anthony B Zwi. 2002. The World Report on Violence and Health. *The Lancet* 360, 9339 (2002), 1083–1088.

[40] Kate Sangwon Lee. 2023. Examining Safety and Inclusive Interventions on Dating Apps by Adopting Responsible Social Media Guidelines. In *Proceedings of the 2023 ACM Conference on Information Technology for Social Good*. 537–546.

[41] David Myles. 2024. Grindr? It'sa "Blackmailer's Goldmine"! The Weaponization of Queer Data Publics Amid the US–China Trade Conflict. *Sexualities* 27, 7 (2024), 1205–1224.

[42] Molly Niesen. 2016. Love, Inc.: Toward Structural Intersectional Analysis of Online Dating Sites and Applications. In *The Intersectional Internet: Race, Sex, Class, and Culture Online*, Safiya Umoja Noble and Brendesha M. Tynes (Eds.). Peter Lang International Academic Publishers, 161–178.

[43] Naheem Noah, Supriya Thakur, Jason Beck, and Sanchari Das. 2024. Evaluating Privacy & Security of Online Dating Applications with a Focus on Older Adults. In *Workshop on Accessible Security and Privacy (WASP 24). IEEE EuroS&P*.

[44] Borke Obada-Obieh, Sonia Chiasson, and Anil Somayaji. 2017. "Don't Break My Heart!": User Security Strategies for Online Dating. In *Workshop on Usable Security (USEC)*. 1–6.

[45] Katarzyna Obarska, Karol Szymczak, Karol Lewczuk, Mateusz Gola, and Mateusz Gola. 2020. Threats to Mental Health Facilitated by Dating Applications Use Among Men Having Sex With Men. *Frontiers in Psychiatry* 11 (2020). https://doi.org/10.3389/FPSYT.2020.584548

[46] Ofcom. 2024. Online Experiences Tracker. https://www.ofcom.org.uk/media-use-and-attitudes/online-habits/internet-users-experience-of-harm-online/

[47] Ofcom. 2024. Online Nation. https://www.ofcom.org.uk/siteassets/resources/documents/research-and-data/online-research/online-nation/2024/online-nation-2024-report.pdf?v=386238

[48] Liliia Pankratova. 2020. Risks of Online Sexual Scripts. *DEStech Transactions on Social Science, Education and Human Science* (2020). https://doi.org/10.12783/DTSSEHS/ICPCS2020/33861

[49] Jessica A Pater, Moon K Kim, Elizabeth D Mynatt, and Casey Fiesler. 2016. Characterizations of Online Harassment: Comparing Policies Across Social Media Platforms. In *Proceedings of the 2016 ACM International Conference on Supporting Group Work*. 369–374.

[50] Kun Peng, Wan Ying Lin, and Hexin Chen. 2022. Consequences of Deceptive Self-Presentation in Online Dating. *Chinese Journal of Communication* 15 (2022). https://doi.org/10.1145/3689710

[51] Olivia Petter. 2018. Why Consent Apps Don't Work, According to Criminal Lawyers. *The Independent* 14 (2018).

[52] Abby R Phillips and Bridget Klest. 2024. Examining the Perceived Harms of Digital Dating Abuse: A University Sample. *Journal of Youth Studies* 27, 1 (2024), 111–124.

[53] Kamarah Pooley and Hayley Boxall. 2020. Mobile Dating Applications and Sexual and Violent Offending. *Trends and Issues in Crime and Criminal Justice* 612 (2020), 1–16.

[54] Urszula Pruchniewska. 2020. "I Like That it's My Choice a Couple Different Times": Gender, Affordances, and User Experience on Bumble Dating. *International Journal of Communication* 14 (2020), 2422–2439.

[55] Kate Raynes-Goldie. 2010. Aliases, Creeping, and Wall Cleaning: Understanding Privacy in the Age of Facebook. *First Monday* (2010).

[56] Belinda Robnett and Cynthia Feliciano. 2011. Patterns of Racial-Ethnic Exclusion by Internet Daters. *Social Forces* 89 (2011). https://doi.org/10.1093/SF/89.3.807

[57] Michael J. Rosenfeld and Reuban J. Thomas. 2012. Searching for a Mate: The Rise of the Internet as a Social Intermediary. *American Sociological Review* 77 (2012). https://doi.org/10.1177/0003122412

[58] Fabio Sabatini and Francesco Sarracino. 2017. Online Networks and Subjective Well-Being. *Kyklos* 70 (2017). https://doi.org/10.1111/KYKL.12145

[59] Morgan Klaus Scheuerman, Jialun Aaron Jiang, Casey Fiesler, and Jed R. Brubaker. 2021. A Framework of Severity for Harmful Content Online. *arXiv: Human-Computer Interaction* (2021). https://doi.org/10.1145/3479512

[60] Charlotte Schluger, Jonathan P Chang, Cristian Danescu-Niculescu-Mizil, and Karen Levy. 2022. Proactive Moderation of Online Discussions: Existing Practices and the Potential for Algorithmic Support. *Proceedings of the ACM on Human-Computer Interaction* 6, CSCW2 (2022), 1–27.

[61] Sarita Schoenebeck, Cliff Lampe, and Penny Trieu. 2023. Online Harassment: Assessing Harms and Remedies. *Social Media+ Society* 9, 1 (2023), 1–13.

[62] Gyanendra Shrestha and Soumya Tejaswi Vadlamani. 2024. Improving the Accessibility of Dating Websites for Individuals with Visual Impairments. arXiv:2410.03695 [cs.HC] https://arxiv.org/abs/2410.03695

[63] Autumn Slaughter and Elana Newman. 2022. New Frontiers: Moving Beyond Cyberbullying to Define Online Harassment. *Journal of Online Trust and Safety* 1, 2 (2022).

[64] Zahra Stardust, Rosalie Gillett, and Kath Albury. 2023. Surveillance Does Not Equal Safety: Police, Data and Consent on Dating Apps. *Crime, Media, Culture* 19, 2 (2023), 274–295.

[65] Tinder. 2024. *About Tinder.* https://tinder.com/en-GB/about

[66] Joris Van Ouytsel, Koen Ponnet, and Michel Walrave. 2018. Cyber Dating Abuse Victimization Among Secondary School Students From a Lifestyle-Routine Activities Theory Perspective. *Journal of Interpersonal Violence* 33, 17 (2018), 2767–2776.

[67] Chenyang Wang. 2022. Online Dating Scam Victims Psychological Impact Analysis. *Journal of Education, Humanities and Social Sciences* 4 (2022). https://doi.org/10.54097/ehss.v4i.2740

[68] Lena Wang. 2020. The Three Harms of Gendered Technology. *Australasian Journal of Information Systems* 24 (2020). https://doi.org/10.3127/AJIS.V24I0.2799

[69] Mark Warner, Agnieszka Kitkowska, Jo Gibbs, Juan F Maestre, and Ann Blandford. 2020. Evaluating 'Prefer Not to Say' Around Sensitive Disclosures. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–13.

[70] Mark Warner, Angelika Strohmayer, Matthew Higgs, and Lynne Coventry. 2025. A Critical Reflection on the Use of Toxicity Detection Algorithms in Proactive Content Moderation Systems. *International Journal of Human-Computer Studies* (2025), 1–15.

[71] Mark Warner, Angelika Strohmayer, Matthew Higgs, Husnain Rafiq, Liying Yang, and Lynne Coventry. 2024. Key to Kindness: Reducing Toxicity in Online Discourse Through Proactive Content Moderation in a Mobile Keyboard. *arXiv preprint arXiv:2401.10627* (2024).

[72] Brenda K. Wiederhold. 2024. Digital Desires, Real Losses: The Complex World of Online Romance Fraud. *Cyberpsychology, Behavior, and Social Networking* (2024). https://doi.org/10.1089/cyber.2024.29311.editorial

[73] Heather Wolbers and Hayley Boxall. 2024. Online Dating App Facilitated Sexual Violence Victimisation Among People With Disability. *Trends and Issues in Crime and Criminal Justice* 695 (2024).

[74] Heather Wolbers and Christopher Dowling. 2024. Routine Online Activities and Vulnerability to Dating App Facilitated Sexual Violence. *Trends and Issues in Crime and Criminal Justice* (2024). https://doi.org/10.52922/ti77697

[75] Sijia Xiao, Shagun Jhaver, and Niloufar Salehi. 2023. Addressing Interpersonal Harm in Online Gaming Communities: The Opportunities and Challenges for a Restorative Justice Approach. *ACM Transactions on Computer-Human Interaction* 30, 6 (2023), 1–36.

[76] Douglas Zytko and Nicholas Furlo. 2023. Online Dating as Context to Design Sexual Consent Technology with Women and LGBTQ+ Stakeholders. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems* (Hamburg, Germany) *(CHI '23)*. Association for Computing Machinery, New York, NY, USA, Article 339, 17 pages. https://doi.org/10.1145/3544548.3580911

[77] Douglas Zytko, Nicholas Furlo, and Hanan Aljasim. 2022. Human-AI Interaction for User Safety in Social Matching Apps: Involving Marginalized Users in Design. *arXiv Preprint arXiv:2204.00691* (2022).

A Appendix: Platform and Document Selection

Platform Names	Monthly Audience	Intended Use of Platforms
Tinder	1 894	Finding love, a date, or just casual conversation [65].
Hinge	1 378	Facilitating people finding love and is "designed to be deleted" [35].
Bumble	1 072	Dating on Bumble challenges heterosexual dating norms by requiring women to make the first move within 24 hours of matching [11].
Grindr	913	Finding friends, dates, casual hook-ups, or relationships for LGBTQ+ individuals, aiming to connect queer people [28][29].
Badoo	521	Chatting, casual dating, or finding relationships [7].

Table 2: This table shows monthly audience use (in thousands) for selected dating platforms, as of May 2024. It also describes the stated aims of the platforms, as described by the platforms.

Platforms	Domains Searched
Badoo	https://badoo.com/
Bumble	https://bumble.com/en-us/help https://bumble.com/guidelines
Grindr	https://help.grindr.com/
Hinge	https://hingeapp.zendesk.com
Tinder	https://policies.tinder.com/

Table 3: Domains used for advanced Google search for each platform.

Search Terms
Harass OR Insult OR Bully OR Embarrass OR Threat OR Doxx OR Troll OR Stalk OR Hate OR Harm OR Griefing OR Unsolicited OR Non-consensual OR Abuse OR Self-harm OR Suicide OR Eating disorder OR Behaviour OR Behavior OR Acceptable OR Community OR Friend requests OR Unwanted OR Unsafe OR Malicious OR Uncomfortable.

Table 4: Domain search terms used for each platform.