RMIT University · Australian Communications Consumer Action Network

Safer by Design

Design and Policy Recommendations for Social Media Safety of Women and Gender-Diverse People in Australia

Authored by Senuri Wijenayake, Madhuka De Silva, Dana McKay, Anastasia Powell, Asangi Jayatilaka, Danula Hettiachchi, Tuck Wah Leong, Joanne E. Gray, Luke Hespanhol, Justine Humphry, Anjalee de Silva

Acknowledgment of Country

Artwork ‘Sentient’ by artist Hollie Johnson

RMIT University and the Australian Communications Consumer Action Network recognise the Traditional Custodians of the lands on which we live, work, and conduct research.

RMIT University acknowledges the people of the Woi wurrung and Boon wurrung language groups of the eastern Kulin Nation on whose unceded lands we conduct the business of the University.
RMIT University respectfully acknowledges their Ancestors and Elders, past and present. RMIT also acknowledges the Traditional Custodians and their Ancestors of the lands and waters across Australia where we conduct our business.

We also acknowledge that the research and community engagement undertaken through this project took place across many parts of Australia, on the lands of many different Traditional Custodians. We extend our respect to all Aboriginal and Torres Strait Islander peoples who are connected to the communities and Country represented in this work.

SAFER BY DESIGN

Design and Policy Recommendations for Social Media Safety for Women and Gender-Diverse People in Australia

Recommended citation

Wijenayake, S., De Silva, M., McKay, D., Powell, A., Jayatilaka, A., Hettiachchi, D., Leong, T.W., Gray, J.E., Hespanhol, L., Humphry, J., & de Silva, A. (2026). Safer by Design: Design and Policy Recommendations for Social Media Safety for Women and Gender-Diverse People in Australia. RMIT University and Australian Communications Consumer Action Network.

ISBN: 978-1-921974-80-9

Funding

This work was funded by the Australian Communications Consumer Action Network (ACCAN). The operation of ACCAN is made possible by funding provided by the Commonwealth of Australia under section 593 of the Telecommunications Act 1997. This funding is recovered from charges on telecommunications carriers.

WEBSITE
https://www.rmit.edu.au/

PHONE
+61 3 9925 2000

POSTAL ADDRESS
RMIT University, GPO Box 2476,
Melbourne VIC 3001 Australia

WEBSITE
https://www.accan.org.au/

PHONE
+61 2 9288 4000

POSTAL ADDRESS
Australian Communications Consumer Action Network, PO Box A1158
Sydney South, NSW 1235

Table of Contents

Acknowledgements

We would like to express our sincere gratitude to all the women and gender-diverse people who generously gave their time to participate in the co-design workshops. Your perspectives, experiences, and ideas are at the heart of this report and have made an invaluable contribution to improving social media safety in Australia.

We would also like to thank the experts who participated in the consultation workshops, who brought a breadth of professional knowledge across platform safety, digital policy, moderation, technology design, and community advocacy.

We acknowledge the Office of the eSafety Commissioner for their participation in the consultation workshops and for their thoughtful input and feedback throughout the project, which strengthened the relevance and impact of this work.

We are sincerely grateful to the Australian GLBTIQ Multicultural Council (AGMC) for their support with participant recruitment, which was essential in ensuring diverse community representation in this project.

We extend our thanks to the Women's Services Network (WESNET) for their ongoing engagement, as well as their valuable insights and feedback that helped shape the direction of the research.

We would like to acknowledge the broader research team members from the University of Sydney and the University of Melbourne, whose contributions strengthened the design and delivery of this study.

This work was funded by the Australian Communications Consumer Action Network (ACCAN).

Terminology

Co-design: A participatory research approach in which those affected by a problem are actively involved in generating and refining solutions.

Culturally and linguistically diverse (CALD): A broad term used to describe communities with diverse languages, ethnic backgrounds, nationalities, traditions, societal structures, and religions.

Deepfake: Realistic images or videos of a person generated or manipulated using artificial intelligence, often without the consent of the person depicted. In the context of this report, deepfakes refer specifically to sexually explicit content created without consent.

Doxxing: The act of publicly revealing private or identifying information about a person online without their consent, typically with the intent to harass, threaten, or harm.

Friction (design): Intentional design features that slow down or interrupt certain user actions, creating a moment of pause before a potentially harmful behaviour occurs.

Gender-diverse: An inclusive term referring to people whose gender identity or expression differs from the sex they were assigned at birth, including but not limited to non-binary, genderqueer, genderfluid, and transgender people.

Image-based abuse: The non-consensual creation, distribution, or threatened distribution of nude or sexual images of a person.

LGBTIQ+: An inclusive term referring to sexuality and gender-diverse communities, encompassing lesbian, gay, bisexual, transgender, intersex, queer, and related identities.

Safety by Design: A framework that encourages technology platforms to proactively embed user safety into the design and governance of their products, rather than addressing harm only after it occurs.

Technology-facilitated abuse (TFA): The use of digital technologies to perpetrate interpersonal harm, including online harassment, stalking, image-based abuse, and coercive control.

Positionality Statement

The research team that designed and conducted this study brings a range of intersecting identities, experiences, and disciplinary perspectives that have shaped how this work was approached and interpreted. The team is made up of women and people from culturally and linguistically diverse backgrounds, with expertise spanning human-computer interaction, digital safety, user experience and design, law, policy, cybersecurity, and qualitative and quantitative methodologies. Several team members have personal experience as social media users navigating safety concerns, and some have direct lived experience of social media abuse. We recognise that these positionalities both enrich and bound our interpretations, and we have sought to centre the voices and ideas of participants throughout the research process.

Executive Summary

This report presents the findings of the Social Media Safety co-design study, a collaboration between RMIT University and the Australian Communications Consumer Action Network (ACCAN). The project draws on 24 structured co-design workshops with 75 women and gender-diverse social media users across Australia, followed by six expert consultation workshops with 21 professionals spanning platform safety, digital policy, moderation, and technology design. The workshops drew on participants' lived experiences and perspectives of using existing platform safety features, exploring how effective these features are against technology-facilitated abuse, their strengths and limitations, and the safer alternatives that users themselves can envision. From these insights, we have compiled six key design and policy recommendations to strengthen how platforms meet the online safety needs and expectations of women and gender-diverse users in Australia.

Online safety is not a peripheral concern. It shapes who can participate fully in digital life. Women and gender-diverse people in Australia disproportionately experience harassment, non-consensual image sharing, impersonation, and identity-based abuse on social media. These harms affect professional visibility, mental health, and everyday communication. Yet existing platform safety tools, from reporting systems to content controls to identity verification, consistently fall short of users' needs, leaving them to manage harm reactively, with little transparency, and at significant personal cost.

> *While technology-facilitated abuse against women and gender-diverse people is recognised as a serious and growing concern in Australia, current platform safety architectures do not adequately reflect the everyday experiences and expectations of those most affected. This report presents six design and policy recommendations, developed directly with users and refined through expert scrutiny; to strengthen how social media platforms prevent, detect, and respond to abuse targeting women and gender-diverse people in Australia.*

Design and Policy Key Recommendations

REC. 1 Make Reporting Transparent, Trackable and Severity-Based

Women and gender-diverse social media users describe reporting as confusing, opaque, and emotionally draining. Experts agree that while scale limits fully human moderation, platforms can improve trust and effectiveness by clearly prioritising urgent harms and communicating what happens after a report is made (see Section 4.1).

DESIGN RECOMMENDATIONS

- Introduce clearer reporting categories with optional fields for additional context to capture situations where harm depends on relationships or identity (harassment from a former partner, impersonation by someone known to the victim, or abuse targeting gender, sexuality, or cultural identity).
- Reporting systems need to be designed to respond appropriately to different types of user reports. Implement hybrid AI-human triage systems that prioritise high-risk cases (intimate image abuse) while maintaining human oversight for nuanced harms.
- Platform reporting systems should provide users with clear and timely information about what happens after a report is made — including timelines, outcomes, and any actions taken. They should also automatically provide status updates to users about their reports.

POLICY RECOMMENDATIONS

- Platform policy can establish clear severity-based escalation standards and clearly articulate how different types of harm are prioritised.
- Platforms can strengthen moderator training in culturally specific and identity-based harms, including non-Western linguistic nuances.
- Platform policy should align reporting processes with emerging digital duty of care expectations (government policies), ensuring procedural fairness and meaningful communication with users.

REC. 2 Embed Friction, Consent, and Evidence-Preserving Controls into Content Sharing

Women and gender-diverse social media users describe experiencing a profound loss of control once intimate or sensitive content begins circulating without their consent. By the time a person notices and submits a report, content has often already spread widely. Platforms can meaningfully reduce this harm by designing platforms for increased user control over the circulation of content (see Section 4.2).

DESIGN RECOMMENDATIONS

- Develop baseline technical capabilities for detecting harmful content, verifying consent, and biometric analysis. These capabilities form the foundation that enables all proactive platform interventions.
- Introduce friction measures such as multiple confirmation prompts to deter thoughtless sharing of harmful content.
- Implement technical measures to prevent users from taking screenshots or downloads of specific content.
- Track and provide users with real-time information about when and where their content is being shared or accessed.
- Automatically log relevant metadata when content is reshared, ensuring records are available for rapid reporting, take down requests and assessing user reports.

POLICY RECOMMENDATIONS

- Government policies and regulatory bodies should develop and enforce data retention, storage, and access standards for platform-generated evidence records, ensuring that verified records are readily available to users seeking to report harm or take legal action, without placing the burden of documentation on those affected.
- Government policies and regulatory bodies should enforce baseline industry standards for detecting and preventing the circulation of harmful or intimate content, verifying consent before intimate or sensitive content involving another person is shared, biometric analysis to support identification of non-consensual image distribution, and transparency obligations requiring platforms to disclose how content detection and moderation decisions are made.

REC. 3 Strengthen Account Authenticity and Accountability Mechanisms

Users express frustration with repeat offenders, fake accounts, and impersonation, particularly when blocking or banning fails to prevent re-entry. Experts acknowledge that stronger identity and account-linking mechanisms may deter casual abuse and support faster resolution of impersonation but emphasise that such measures must be carefully balanced against privacy, safety, and exclusion risks (see Section 4.3).

DESIGN RECOMMENDATIONS

- Introduce multiple verification options (phone number, email, device linking, or optional ID checks) to add friction to repeat account creation without requiring all users to disclose their real identity.
- Introduce additional verification when high-risk behaviours are detected (repeated bans, impersonation reports, or coordinated harassment), rather than requiring verification for all users.
- Provide clearer signals of account authenticity (verified identity markers or account history indicators) to help users make informed decisions about engagement.
- Allow flexible multi-account management while monitoring abuse patterns across linked accounts.

POLICY RECOMMENDATIONS

- Ensure platform policies governing verification systems require collecting only the minimum personal information needed, storing it securely, setting clear limits on how long it is kept, and transparently explaining how identity information is handled.
- Ensure platform policies avoid mandatory identity verification that may exclude users without formal ID, those in unsafe domestic situations, or individuals who rely on pseudonymity for protection.
- Align platform policies on identity verification with relevant government regulations across jurisdictions, including safeguards against identity tracking or surveillance.

REC. 4 Strengthen Context-Sensitive Flagging and Detection Mechanisms

Users want greater ability to flag harmful content, not only as victims but also as bystanders. They also express frustration when repeated re-uploads and culturally specific abuse go undetected. Experts support improved flagging and media-matching systems to reduce the burden on affected users, but caution that definitions of 'harm' are context-dependent and that automated detection alone cannot reliably interpret cultural nuance or evolving language (see Section 4.4).

DESIGN RECOMMENDATIONS

- Enable bystander flagging pathways that include clear guidance on what constitutes harmful content, how flags will be weighted and assessed, and how the person affected will be notified in a trauma-aware manner.
- Incorporate AI-assisted media matching to identify identical or near-identical re-uploads of previously flagged content, reducing the need for repeated victim reporting.
- Combine automated detection with human oversight in culturally sensitive or context-dependent cases.
- Provide configurable notifications when content involving a user is flagged or detected, ensuring such alerts are trauma-aware and optional to avoid re-triggering harm.
- Regularly review and update detection models to account for evolving slang, coded language, and culturally specific expressions of abuse.

POLICY RECOMMENDATIONS

- Establish platform policies governing flagged-content databases that align with relevant government regulations and define strict controls on how sensitive media is stored, accessed, and retained to prevent redistribution or secondary harm.
- Develop platform policies that include safeguards against misuse of flagging systems, including protections against coordinated or malicious reporting.
- Ensure moderation policies recognise cultural and linguistic variation, with investment in culturally competent review processes across jurisdictions.
- Maintain transparency within platform policies about how flagged content is evaluated, while avoiding disclosure that could enable evasion of detection systems.

REC. 5 Improve Detection and Accountability for Repeat and Patterned Abuse

Users want platforms to recognise patterns of repeated abuse, including multiple reports against the same account, repeated stalking-like behaviour, and suspicious activity across accounts. Many express a desire for greater visibility into potential risk signals (repeated profile visits or linked accounts), particularly in situations involving known perpetrators. Experts acknowledge the potential value of pattern detection in reducing harm and supporting earlier intervention, but emphasise that numerical indicators alone can be misleading and must be contextualised to avoid unfair targeting or misuse, including the disproportionate flagging of marginalised users or the weaponisation of reporting systems against the very people they are designed to protect (see Section 4.5).

DESIGN RECOMMENDATIONS

- Connect reports over time to detect repeat patterns of harassment or abuse, rather than treating incidents as isolated events.
- Develop internal risk signals (repeated account creation, coordinated targeting, linked identifiers) to support earlier moderation intervention.
- Provide carefully designed user-facing safety signals (alerts about unusual or repetitive activity), prioritising user control and configurability.
- Ensure that any visibility features (profile view indicators or linked-account detection) are opt-in and balanced against privacy risks.

POLICY RECOMMENDATIONS

- Establish platform policies about process standards for pattern-based enforcement, including thresholds for action, documentation of decisions, and accessible appeal pathways.
- Implement platform safeguards against false reporting and misuse of flagging.
- Define the platform's limits on what behavioural data is surfaced to users to prevent retaliation, doxxing, or escalation.
- Ensure that the platform's pattern detection systems are audited for bias, particularly where marginalised communities may be disproportionately reported or misclassified.

REC. 6 Strengthen Safety Awareness Through Contextual and Responsible Guidance

Users describe difficulty discovering or understanding existing safety tools, often only encountering them after harm had escalated. They want clearer, timely guidance about what features exist and when to use them. Experts agree that improving discoverability and contextual support could reduce harm and user burden (see Section 4.6).

DESIGN RECOMMENDATIONS

- Improve discoverability of safety features by offering contextual guidance to users (suggesting blocking, reporting, or relevant privacy controls when harmful interactions occur or are predicted), with AI-assisted guidance as one possible mechanism for surfacing these features when needed.
- Apply progressive disclosure: stage safety education over time instead of front-loading dense onboarding flows that users are likely to skip.

POLICY RECOMMENDATIONS

- Establish platform guardrails for AI-based safety mechanisms, ensuring they redirect users to verified resources rather than independently interpreting or advising on high-risk situations.
- Platforms should define clear protocols for how sensitive information shared by users with AI-based safety assistants is logged, stored, and escalated, particularly where users disclose high-risk situations such as coercive control, threats, or self-harm.

Section 1

Introduction

This report presents findings from a national co-design study examining how women and gender-diverse people in Australia engage with social media safety features in the context of technology-facilitated abuse (TFA). The objective of this research was to build a new evidence base centred on everyday interactions with platform safety tools, and to translate these insights into practical design and policy recommendations for strengthening social media safety. These findings are intended to help designers and platform developers envision improved and alternative safety features to those that currently exist, by identifying what works, what falls short, and where gaps remain, particularly for women and gender-diverse users whose needs and expectations of safety are not yet adequately reflected in platform design.

Technology-facilitated abuse has become an increasingly visible and harmful dimension of online life, particularly on social media platforms where everyday communication, public expression, and community participation occur. These environments can also create new opportunities for harassment, intimidation, non-consensual image sharing, and identity-based abuse [7, 11, 12, 13]. Women and gender-diverse people are disproportionately affected by these harms, with women in Australia significantly more likely than men to report fearing for their safety online [13], and one in five women aged 18 to 45 reporting the non-consensual sharing of private or intimate content [17]. These insights highlight the importance of examining how safety features function within social media platforms [15, 18], where a substantial proportion of technology-facilitated abuse occurs [11, 33], and where women and gender-diverse users face heightened risk of harm.

1 in 5

women aged 18–45 experienced non-consensual sharing of private content in Australia

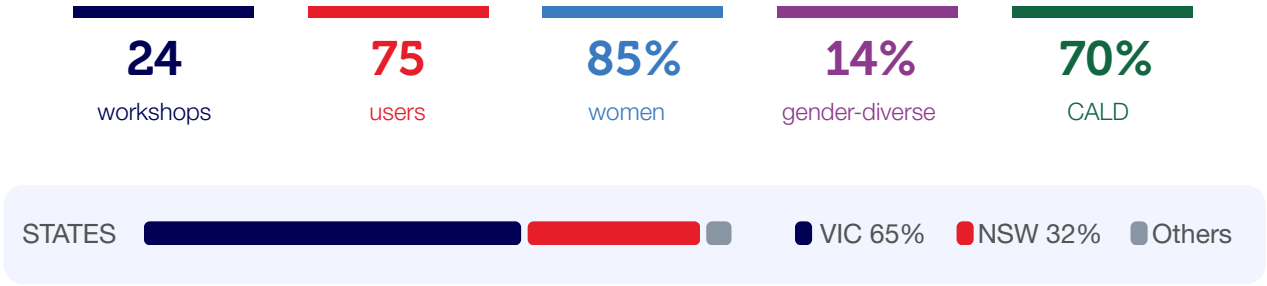
Women are significantly

more likely than men to report fearing for their safety online.

Despite the ongoing expansion of social media safety architectures, limited Australian research has examined how adult women and gender-diverse users experience and navigate these tools in practice. Much existing work documents the prevalence and harms of technology-facilitated abuse [7, 11, 13, 14, 24]. However, less is known about how privacy and safety features — including but not limited to reporting systems, blocking functions, content controls, and identity verification mechanisms — operate in everyday use, where they fall short, and how users themselves imagine safer alternatives.

To address this gap, we conducted 24 structured co-design workshops with 75 social media users across Australia, including women (85%) and gender-diverse participants (14%), with substantial representation from culturally and linguistically diverse (CALD) communities (70%). Participants were based across several Australian states, with most living in Victoria (VIC) (65%), followed by New South Wales (NSW) (32%), and few from South Australia and the Australian Capital Territory. The workshops explored how participants interact with existing safety and

privacy features, alongside the pain points, workarounds, and unmet needs that shape their decision-making. Following the user workshops, we engaged experts in platform safety, digital policy, moderation, and technology design to examine feasibility, governance implications, and potential risks associated with the directions suggested by users.



The report begins with a background section outlining the landscape of technology-facilitated abuse and existing platform responses. It then describes the co-design methodology and expert consultation process. The findings section presents key themes emerging from the workshops, including reporting processes, content sharing controls, identity and verification mechanisms, flagging and detection systems, tracking abusive behaviours, and improving safety awareness and feature discoverability. The discussion integrates these themes to reflect on the definition of harm, burden on the victim, transparency for trust, design friction for prevention, AI influence and systemic changes.

This study contributes new evidence base on how social media safety features are experienced and used in everyday practice. Drawing on insights from the co-design workshops and expert consultations, we identify six key recommendations aimed at strengthening both the design and governance of platform safety systems. These recommendations address areas such as improving transparency in reporting processes, introducing preventative design measures that reduce opportunities for abuse, making safety tools easier to discover and use, and strengthening accountability in how platforms respond to harm.

6
RECOMMENDATIONS

Transparent reporting · preventative design · safety feature discoverability · platform accountability · identity verification · safety awareness

Section 2

Background

Women in Australia experience disproportionate levels of online abuse across social media platforms. National research shows that women are twice as likely as men to fear for their safety online [13,14]. Online abuse directed at women commonly includes harassment, threats, sexualised comments, stalking, and impersonation. Image-based abuse represents one particularly harmful form of this broader pattern. National surveys have found that 20% of women aged 18 to 45 years have experienced non-consensual sharing of private content [7]. The prevalence is higher among respondents who identified as lesbian, gay, or bisexual (19%) compared to those who identified as heterosexual (11%) [7]. These figures demonstrate the scale and gendered nature of technology-facilitated abuse within Australia's social media environment.

2 x

women are twice as likely as men to fear for their safety online

1 in 4

women hesitant to pursue public-facing roles due to online harassment

Beyond exposure rates, research examining the impact of online abuse on women's working lives has found that one in four women reported hesitancy to pursue public-facing roles due to concerns about online harassment [11]. These findings demonstrate that social media abuse extends beyond individual incidents, shaping professional participation, visibility, and digital inclusion.

20%

experienced non-consensual sharing of private or intimate content

Cultural context further shapes experiences of abuse. Australia is one of the most culturally diverse nations in the world, with recent census data showing that around 27.6% of the population were born overseas and nearly half (48.2%) have at least one parent born overseas at the time of the 2021 Census [3]. 5.8 million people (22.8% of the population) reportedly speak a language other than English at home, reflecting linguistic diversity that spans hundreds of languages [4]. These figures underscore the multicultural fabric of Australian society and highlight the significance of considering cultural and language diversity when examining online experiences, including safety and abuse. Research commissioned by the eSafety Commissioner with women from culturally and linguistically diverse (CALD) backgrounds points to additional dimensions of online abuse, such as culturally specific threats, community-related stigma, and

language barriers that inhibit access to support services [24]. These factors may compound harm and create unique barriers to navigating safety features on social media platforms, suggesting that one-size-fits-all approaches to platform safety can miss important contextual needs.

27.6%
born overseas

48.2%
have a parent born overseas

22.8%
speak a language other than English at home

Research shows that online abuse does not occur in isolation from broader social dynamics. Women, gender-diverse people, and those from culturally and linguistically diverse communities often experience abuse that draws directly on gendered, racialised, or identity-based stereotypes. For example, studies of transgender and non-binary users demonstrate how platform features designed for visibility and engagement can inadvertently expose them to targeted harassment or unwanted attention, and how these affordances can be weaponised against them [29]. Research on non-consensual image distribution further illustrates how perpetrators exploit anonymity, virality, and platform affordances to amplify harm [5]. These findings highlight that abuse is shaped not only by individual behaviour but also by how platform tools operate in social contexts where existing power imbalances can be reproduced and intensified. As a result, the same safety feature may function differently depending on who is using it and under what circumstances.

While a growing body of work examines abuse on social media platforms affecting women and gender-diverse people, existing research tends to focus on prevalence, legal reform, or victim experiences of harm. In parallel, fewer studies have used participatory and co-design approaches to explore how social media platforms might better respond to online harassment. These studies often engage participants in scenario-based workshops where harassment situations are discussed through short narratives or storyboards, allowing participants to reflect on current platform responses and collaboratively generate ideas for interventions [1]. Through these activities, participants have proposed a range of design responses, including improvements to reporting systems, visibility controls, and other preventative mechanisms aimed at interrupting harmful interactions or supporting safer platform use. Further research has involved young people in examining how algorithms and platform design shape privacy and safety experiences through participatory design activities, including prompt-based discussions about platform features and algorithms and collaborative sketching exercises using tools such as Figma to surface ideas that help users better understand or manage platform behaviours [26]. Through these approaches, co-design research has highlighted the potential of participatory methods to generate design ideas that address both preventative and responsive forms of online safety.

However, participatory and co-design research in this area has more frequently centred adolescents or youth populations [1, 26], often outside the Australian context. There remains comparatively limited research examining, in depth, how adult women and gender-diverse users in Australia engage with the safety features already embedded within major social media platforms. In particular, there is limited qualitative evidence exploring what tools users rely on, which features are perceived as ineffective or burdensome, and what gaps remain between platform affordances and lived digital realities.

The eSafety Commissioner's national survey found that most adults from targeted groups (including sexually diverse people, Aboriginal and/or Torres Strait Islander people, people with disability, and linguistically diverse communities) did not take action after encountering online hate; among those who did, the most common responses were blocking or reporting the account [12]. Research with women from culturally and linguistically diverse backgrounds similarly indicates that reporting can be perceived as burdensome or ineffective in addressing harm [24]. At the same time, platforms have continued to expand their safety infrastructures, incorporating automated moderation, AI-driven detection, filtering tools, and layered privacy controls. This evolving landscape raises important questions about whether these developments meaningfully address previously identified concerns, why certain features remain underused or perceived as ineffective, and how users themselves imagine more responsive and preventative approaches to online safety.

This report contributes new and in-depth evidence drawn from a substantial program of co-design workshops with adult women and gender-diverse social media users across Australia, including strong representation from culturally and linguistically diverse communities. The scale of participation exceeds many existing qualitative studies in this domain, enabling a broader range of scenarios, perspectives, and safety concerns to be surfaced. Importantly, the study also integrates expert workshops involving professionals in platform safety, digital policy, moderation, and technology governance. This dual-perspective approach allows for both user-generated proposals and critical examination of feasibility, proportionality, governance implications, and unintended consequences.

Section 3

Methodology

To examine social media safety practices and identify opportunities for improvement, we conducted two rounds of co-design workshops with two distinct participant groups: women and gender diverse social media users, and domain experts (online safety, usability and user experience, legal, policymakers). The workshops were designed to support open discussion, collaborative exploration, and the generation of user- and practice-informed insights to inform design and policy recommendations.

Between October and December 2025, 24 structured workshops of approximately three hours were conducted with social media users across Australia, using Zoom and in person sessions (in Melbourne). Subsequently, 6 workshops were conducted in December 2025 with Australian and international experts. Ethical approval was sought and received from the RMIT University Human Research Ethics Committee prior to commencing the research work.

Participant Profiles and Recruitment

Social Media Users

The first round of workshops involved social media users and focused on understanding everyday safety concerns, perceptions of existing platform features, and ideas for improvement. Across this phase, we conducted 24 workshops with a total of 75 participants. Recruitment was carried out through targeted outreach, community networks, and partner organisations, with the aim of engaging a diverse range of social media users. Recruitment materials outlined the purpose of the workshops, participation requirements, and data use practices, and interested participants registered via an online form. As part of this form, participants were asked to provide information about how they used social media, including the platforms they used and the nature of their engagement, to support participant selection and contextualise their perspectives during analysis.

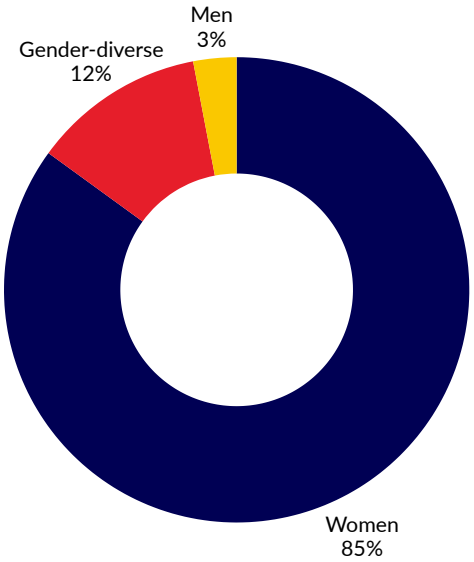


Figure 1. Gender of Users

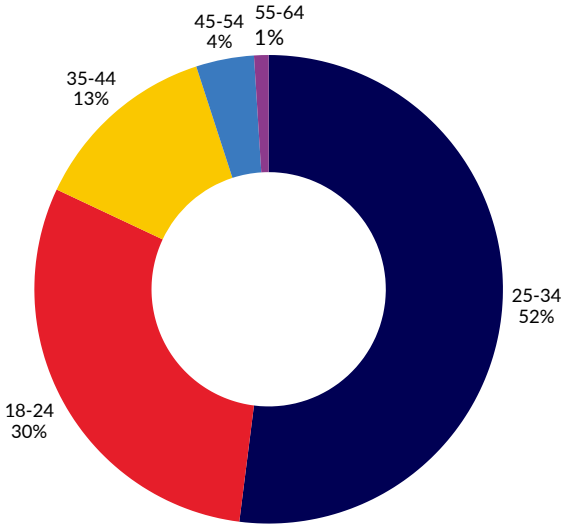


Figure 2. Age ranges of Users

As required by the target group of this research, most participants were women (85%). Eight percent identified as non-binary. One participant (1.3%) identified as a transgender woman, and two (2.6%) identified as genderqueer or genderfluid. Less than three percent (n=2) of participants identified as men whose sexual orientation is gay. The majority (68%) of the sample identified as heterosexual. Thirteen percent identified as bisexual. Ten percent identified as queer, with the remaining participants identifying as asexual (2%), lesbian (1%), and pansexual (1%). Participants ranged from 19 to 60 years old. The average age was 29 years.

Most participants identified as Asian (78%), followed by European (17%), with the remaining identifying as Australian (5.3%) or Middle Eastern (2%). Most participants spoke a language other than English at home (76%), while (18%) reported speaking only English. The languages spoken other than English included Mandarin (18%), Indonesian (18%), Sinhala (7%), Hindi (7%), Cantonese (6%), Tamil (3%), Urdu (3%), Japanese (1%), Kannada (1%), Nepali (1%), Tagalog (1%), Arabic (1%), German (1%), and French (1%).

Most participants were living in Victoria (65%), including one person in a regional area, followed by New South Wales (32%), with the remaining participants residing in South Australia (1%) and the ACT (1%).

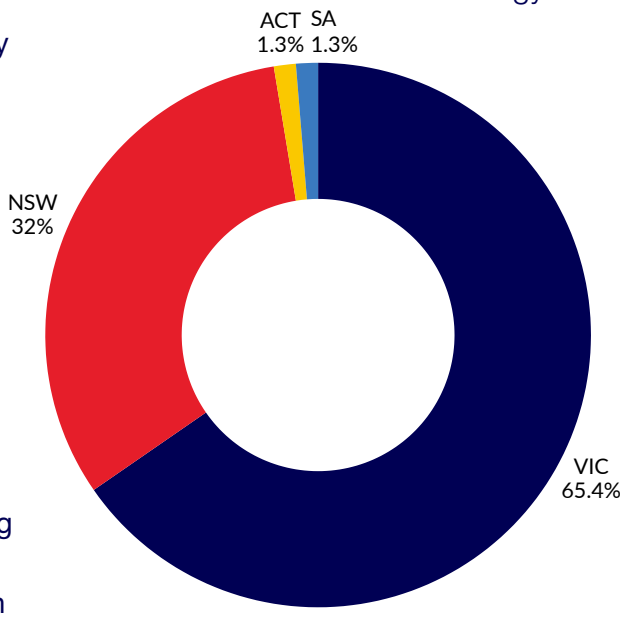


Figure 3. Location of Users

Experts

The second round of workshops was conducted with experts working in areas related to protecting women and gender-diverse individuals from technology-facilitated abuse. This included researchers, industry stakeholders, community advocates, and policymakers. In total, six expert workshops were conducted with 21 participants. These workshops built directly on insights from the user workshops and focused on critique, refinement, and evaluation of the design and policy ideas generated by users.

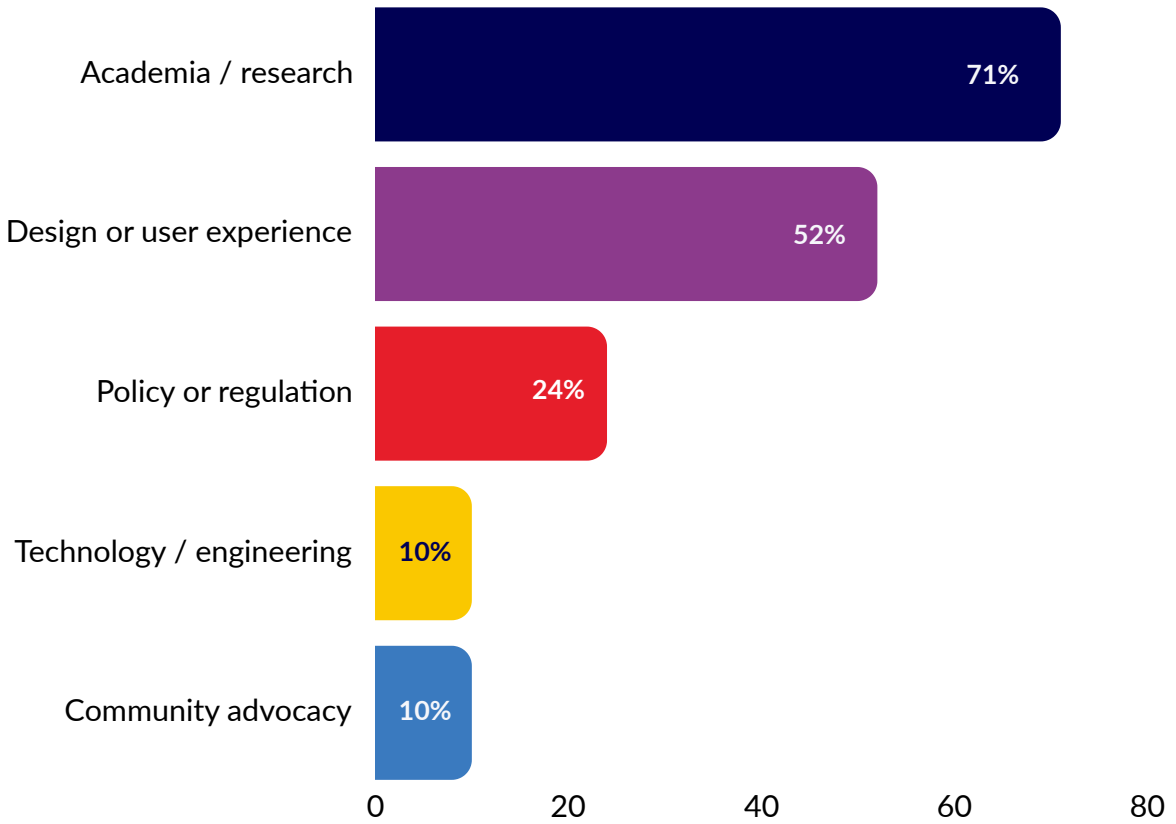


Figure 4. Background of Experts

Experts participating in the consultation workshops brought a range of professional perspectives relevant to online safety, platform governance, and technology design. Most participants had backgrounds in academia or research (71%), alongside expertise in design or user experience (52%), policy or regulatory roles (24%), technology or engineering (10%) and community advocacy or support services (10%). On average, participants had approximately ten years of professional experience in their respective fields. A further table in the appendix provides additional details on the professional experience and areas of expertise represented among the experts.

Participants were primarily based in Australia, with the largest representation from Victoria (43%) and New South Wales (38%), followed by the Australian Capital Territory (10%) and Queensland (5%). One participant was based in the United States (5%), providing an additional international perspective.

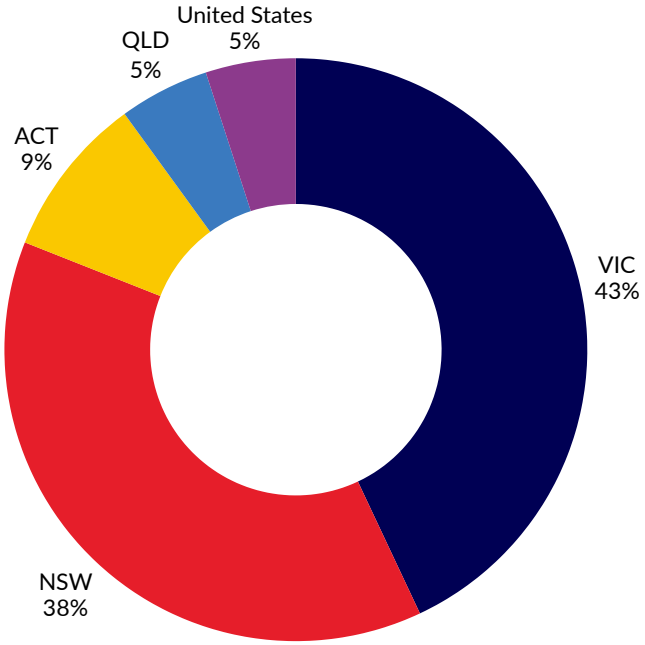


Figure 5. Location of Experts

Experts represented diverse gender identities and cultural backgrounds. Gender distribution included 67% women and 33% men. Participants also reflected a range of ethnic backgrounds, including North-West European (48%), Southern and Eastern European (24%), Oceanian (24%), Southern and Central Asian (14%), North African and Middle Eastern (14%), South-East Asian (10%), as well as smaller representation from Sub-Saharan African, North-East Asian, and American backgrounds. Many participants reported multilingual abilities, with 48% speaking languages in addition to English. Languages spoken included Hebrew, Thai, Sinhala, Hindi, Welsh, Finnish, Spanish, Portuguese, and Arabic, alongside English.

Structure of Workshops with Social Media Users

Co-design workshops were selected as the primary method to engage participants directly, listen to their perspectives, and understand how they interpret and respond to safety challenges on social media. User workshops explored what safety problems participants encounter, how effective or efficient current platform features are perceived to be, and what new or improved approaches they would propose. Expert workshops then engaged participants in critically reflecting on these user-generated ideas, drawing on their professional knowledge to assess feasibility, risks, and broader implications.

Methods

Each workshop was facilitated by two members of the research team and followed a predetermined structure to ensure consistency across sessions. Workshops began with an overview of the goals and scope of the session, followed by a brief introduction to existing social media safety features. We focused specifically on three social media platforms: Facebook, Instagram, and TikTok. These platforms were selected because they are among the most widely used social media platforms in Australia [16], they share similar social media affordances making design and policy recommendations applicable across all three, and they are platforms where a

significant proportion of technology-facilitated abuse has been reported [34, 35]. A short onboarding activity was used to familiarise participants with the collaborative tools being used, particularly the digital whiteboard. Participants then completed a brief icebreaker activity focused on their general use of social media.

The core workshop activities were structured around a scenario-based approach (see next sub-section). Participants were first introduced to a scenario relevant to online safety and engaged in an empathy mapping activity to reflect on how they might think, feel, and act in that situation (see Figures 7 and 8). This activity prompted participants to consider their emotional responses, what questions they might ask, and what actions they or others — including friends, family, and the platform — might take in response to the scenario.

Following the empathy mapping activity, participants developed an individual action plan outlining the steps they would take if they were directly affected by the scenario, drawing on existing platform safety features or other responses available to them (see Figures 9 and 10). Participants then shared and discussed their action plans with the group, offering feedback and identifying common themes across their responses.

Building on this discussion, participants individually developed user stories to articulate specific safety features or improvements they would like to see on social media platforms. These user stories followed the format: 'As a [user], I want [feature], so that [goal]'. The group then voted on the user stories, with each participant selecting one to two stories they felt were most important or feasible to take forward.

Selected user stories formed the basis of a collaborative storyboarding activity, in which participants visually narrated how an improved safety feature might look and function in practice (see Figures 11 and 12). Storyboards were designed to capture both preventative and responsive dimensions of safety — that is, how a feature might help detect or interrupt harm before it escalates, as well as how it might support users in responding after harm has occurred. Participants were prompted to consider questions such as: How does the situation begin? How does the feature intervene or support the user? How is the situation resolved, and does the user feel safe?

Figure 6: Workshop Flow

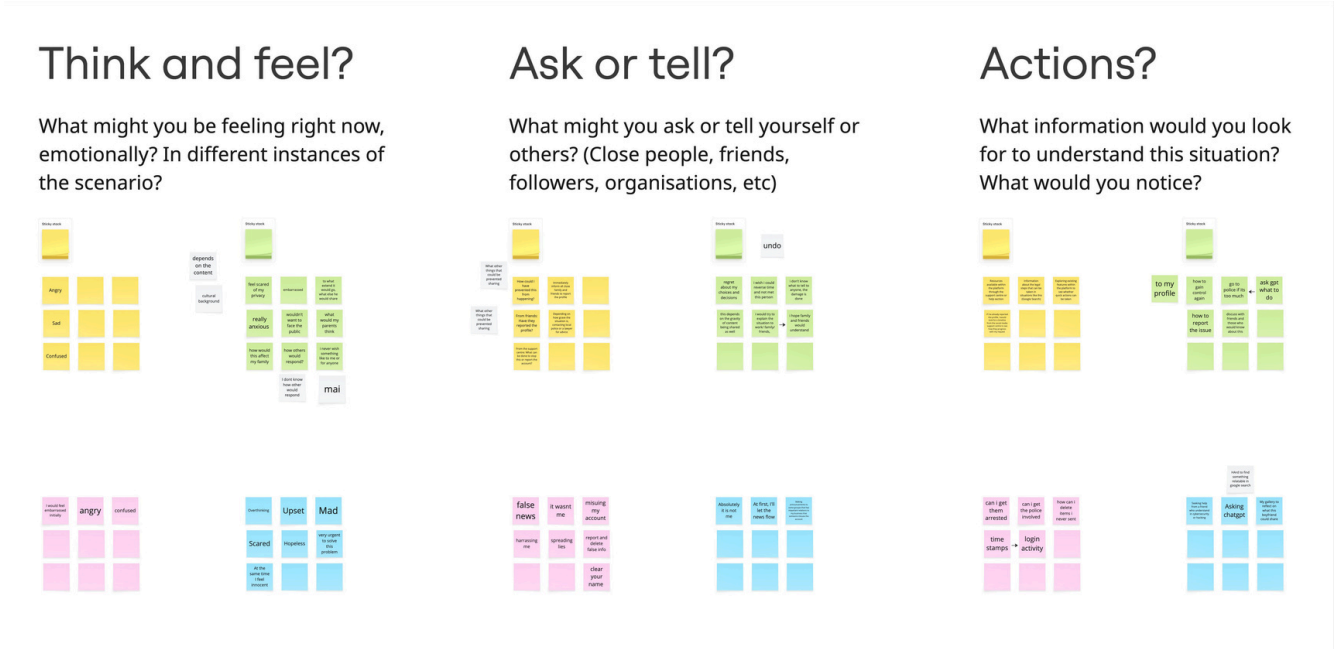


Figure 7: Example of the empathy map activity completed by participants in an online workshop, capturing how they might think, feel, ask, and act in response to a given scenario.

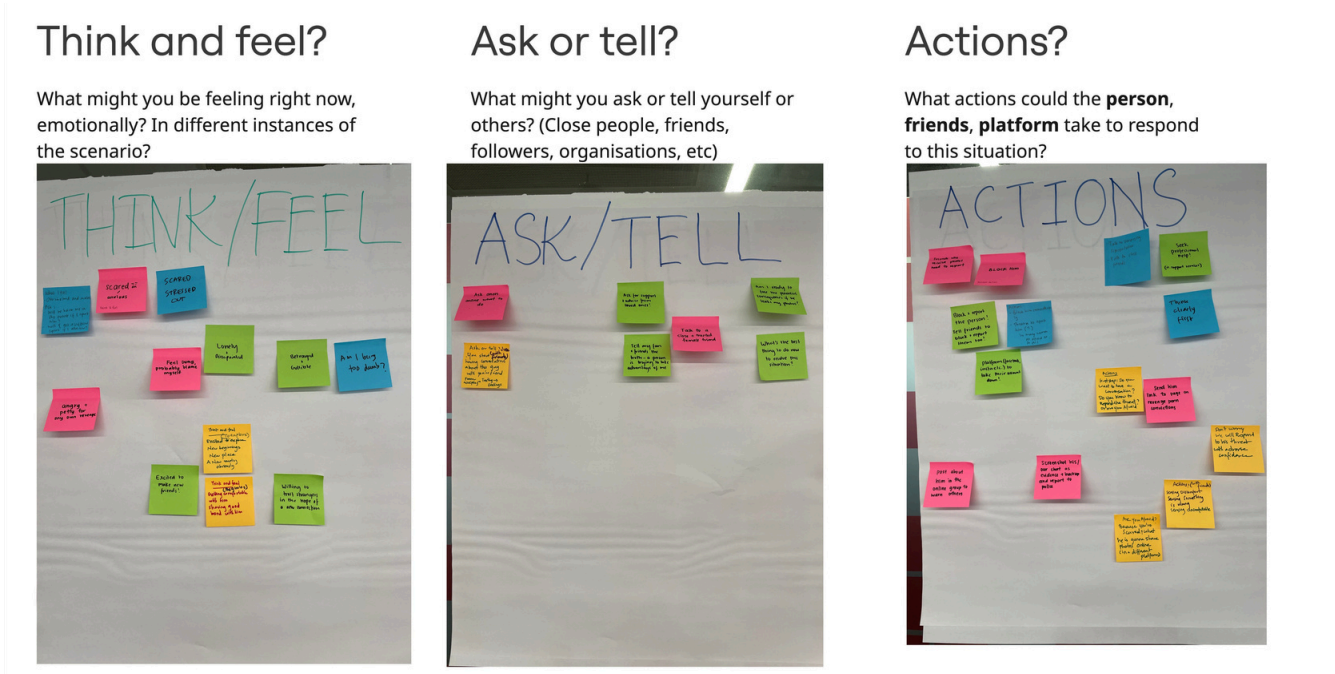


Figure 8: Example of the empathy map activity completed by participants in an in-person workshop, with sticky notes capturing emotional responses, questions, and proposed actions in relation to a given scenario.

Feature Name of the feature you will use or not use	Use & Reason How you would use it and why (**respond vs prevent**, effectiveness)	Knew Before? Yes/ No	Resolved? Improvements Will it solve the problem fully? Suggested new features or improvements	
Report account	respond: it's specifically for impersonation claims	Y	won't solve the problem fully because the account will be frozen, but at least it will stop further damage	SUGGEST: facial/vocal verification
Manage post audience/ visibility OR profile locking	respond: if she really can recover her account, it may benefit her to adjust the privacy of her posts in the interim while she deletes the damage he's done	Y	won't solve the problem fully, only curb further damage while she fixes things	but it can stop her content from being reshared, esp in a way that's traceable back to her

Figure 9: Example of the feature evaluation activity completed by participants in an online workshop, assessing existing platform safety features in terms of their purpose, prior awareness, and potential for improvement.

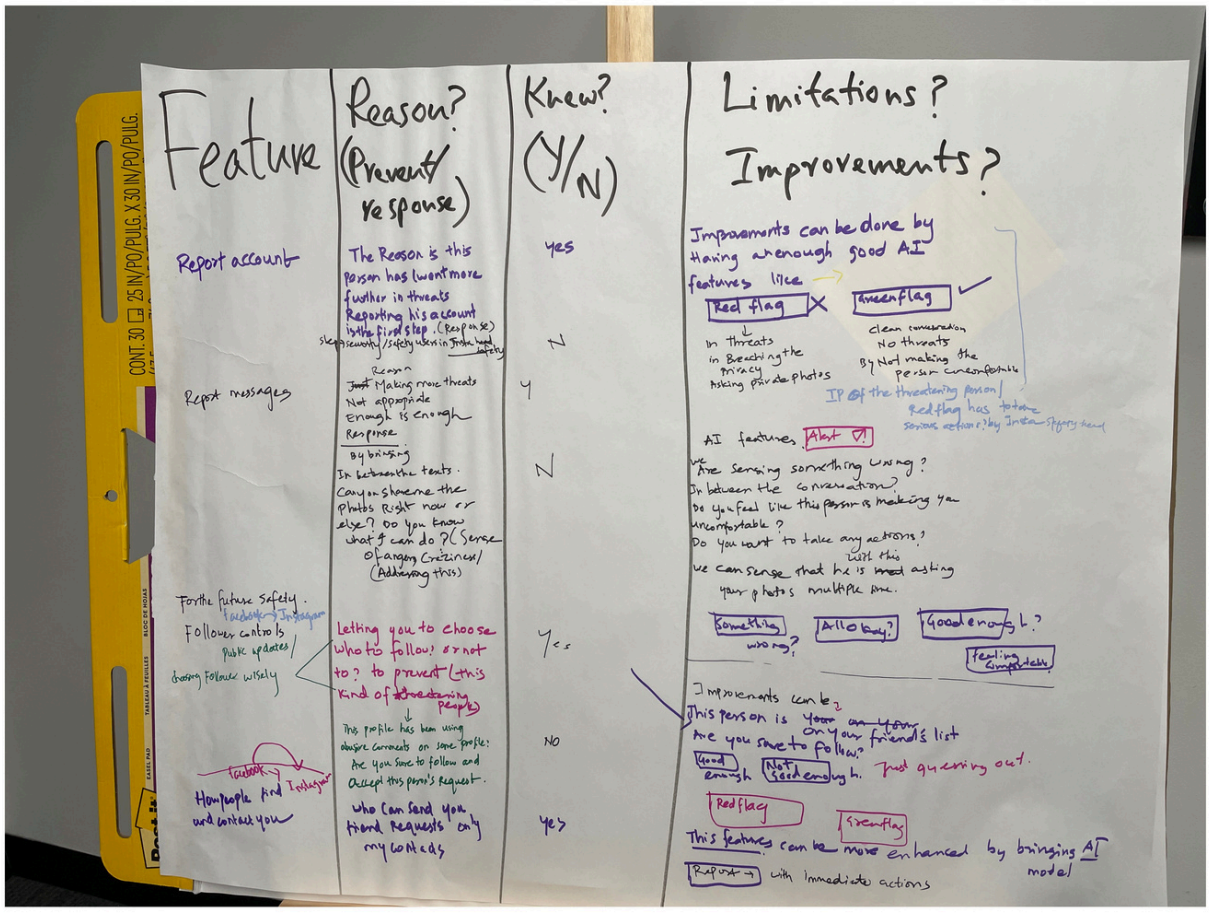


Figure 10: Example of the feature evaluation activity completed by participants in an in-person workshop, with handwritten notes evaluating platform features by reason for use, prior knowledge, and suggested limitations and improvements.

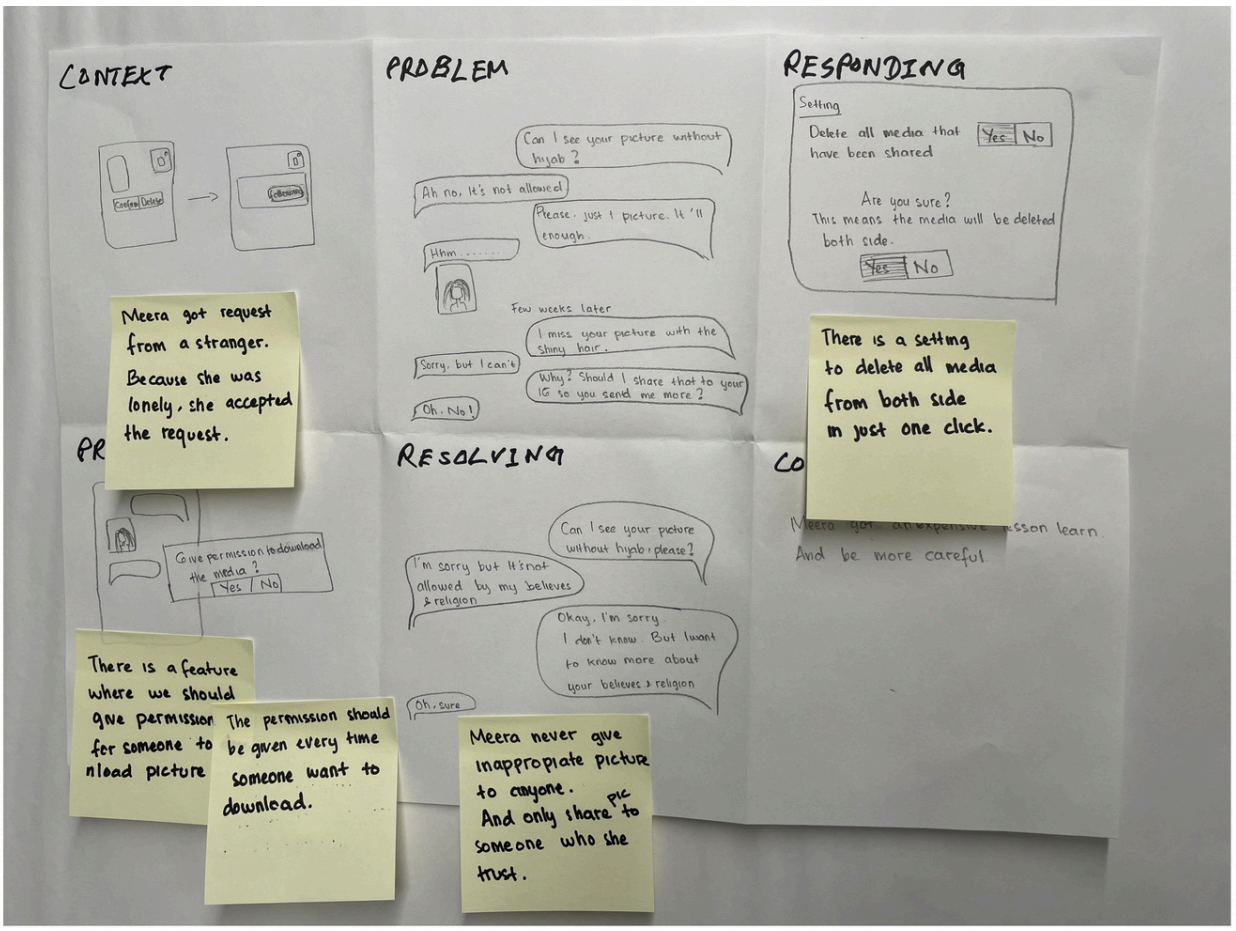


Figure 11: A participant-created storyboard from an in-person workshop illustrating a proposed platform response to a scenario involving unwanted image requests, including a feature to delete shared media from both sides of a conversation.

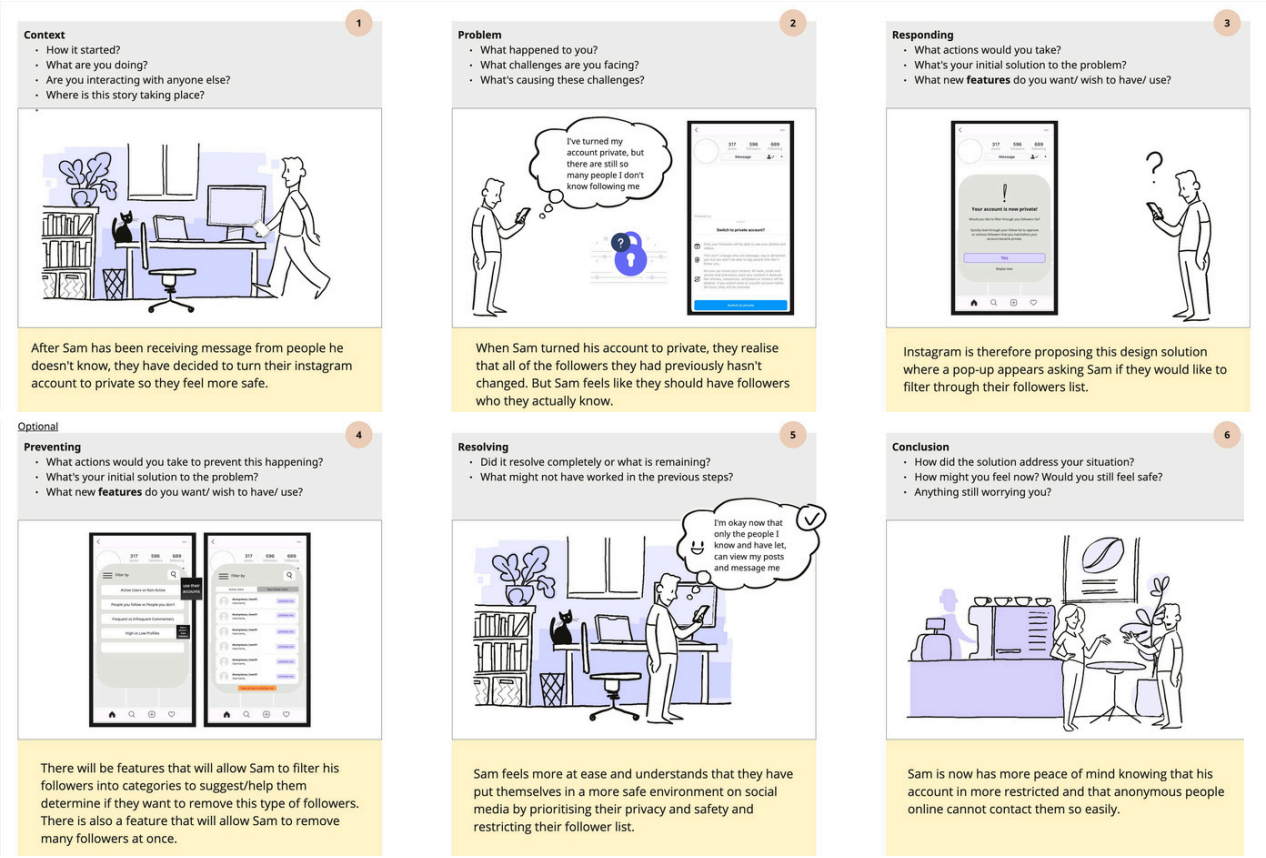


Figure 12: A participant-created storyboard from an online workshop proposing a follower-filtering feature.

Scenarios

To support structured discussion during the workshops, we developed 12 short scenarios depicting different forms of technology-facilitated abuse occurring on social media platforms. These scenarios were designed to reflect common, yet varied, experiences of online harm, enabling participants to respond to concrete situations rather than abstract questions. Each scenario was presented as a short narrative involving a fictional character navigating a specific safety concern. Across the 24 workshops, each scenario was used twice, ensuring that all 12 scenarios were discussed by multiple groups.

The scenarios were informed by prior empirical research and national reports on technology-facilitated abuse, including qualitative interviews with victim-survivors of image-based abuse, research on non-consensual image distribution and deepfake sexual imagery, and reports examining stalking, impersonation, coercive control, and harassment in digital contexts [12, 15, 18, 21, 27]. For example, scenarios involving threats to share intimate images and digitally altered sexual imagery were inspired by research documenting the harms of image-based abuse [21], while situations depicting coercive monitoring, impersonation, and repeated unwanted contact were informed by interview-based studies of victim-survivors and research on technology-facilitated domestic abuse [15,18]. Identity-based harassment and online hate scenarios drew on national research examining targeted abuse against culturally diverse, gender-diverse, and other marginalised communities [12].

Together, the scenarios were designed to capture a spectrum of harms, including stranger harassment, abuse by former partners, impersonation, account compromise, AI-generated explicit imagery, coordinated harassment, and threats of exposure. They also varied by victim profile, perpetrator profile, and relationship dynamics to reflect the relational and contextual nature of social media abuse. Table 1 summarises the 12 scenarios, including the type of abuse, victim and perpetrator characteristics, relationship between them, and the research sources informing each case. The full scenario narratives are provided in Appendix - Scenarios at the end of this report. The goal was not to replicate any single case, but to create realistic prompts that could elicit participants' reflections on existing platform features, perceived limitations, and desired improvements.

Participant Inclusion and Model of Care

Participant inclusion and care were embedded as core principles in the design and facilitation of the workshops. Prior to each session, participants were provided with detailed information about the workshop structure, activities, and expectations to support informed participation. A brief preparatory call was conducted with each participant to understand their motivations for participating, identify accessibility needs, and discuss preferred modes of engagement. Multiple participation options were supported, including verbal discussion, written responses in shared documents, use of chat functions, and contributions via digital whiteboards, enabling participants to engage in ways that felt comfortable and safe.

Workshops were conducted both online and in person, with the format selected based on participant accessibility and location. Participants based interstate or requiring travel were prioritised for remote participation, while in-person sessions were offered where this was more

Table 1 Summary of the 12 scenarios (see assigned reference letter) used across the workshops, including the type of abuse, victim and perpetrator profiles, their relationship, and the research informing each case.

	Type of abuse	Victim's Profile	Perp's Profile	Relation	Source
A	Sending sexual content	Young woman	Middle-aged man	Stranger	[27]
B	Impersonation	Middle-aged gender-diverse	Young woman	Work Colleague	[11]
C	Threatening to post nudes online	Young CALD woman	Young man	Stranger	[15]
D	Posting negatively about an ex-partner	Middle-aged woman	Middle-aged man	Ex-partner	[15]
E	Gender-identity-targeted abuse	Young sexually diverse person	Young man	Friend	[15]
F	Hacking accounts	Young woman	Young man	Ex-partner	[15]
G	Non-consensual intimate images sharing	Young woman	Young man	Current partner	[21]
H	Sending threatening messages	Middle-aged woman	Middle-aged man	Ex-partner	[15]
I	Deepfake pornography	Young woman	Middle-aged man	Stranger	[21]
J	Continually engage with victim's online content	Young woman	Young man	Ex-partner	[15]
K	Cultural-identity-targeted abuse	Middle-aged woman	Any aged person	Stranger	[12]
L	Stalking	Young non-binary person	Young man	Stranger	[12]

accessible. In-person sessions included snacks and refreshments. Across both formats, participants were free to step out at any time (with notice to facilitators), and regular breaks were built into the workshop structure to reduce fatigue and emotional strain.

Given the sensitive nature of the scenarios discussed, a structured model of care was implemented. All facilitators completed formal training in trauma-informed research practice prior to conducting workshops with participants. Participants were invited to anonymously indicate their stress levels throughout the session and were encouraged to signal distress or request a pause using both verbal and anonymous mechanisms. Where distress was indicated, facilitators paused the workshop and guided participants through brief grounding or de-escalation activities drawn from trauma-informed facilitation practice. These included box breathing, in which participants were guided through slow, equal-count inhale, hold, exhale, and hold cycles to regulate the nervous system, and a grounding exercise using the five senses, in which participants were invited to notice five things they could see, four they could hear, three they could touch, two they could smell, and one they could taste, to support a return to present-moment awareness.

Structure of Workshops with Experts

In the expert workshops (Figures 13 and 14), we presented synthesised findings from the user co-design sessions, tailored to align with the specific expertise and professional backgrounds of the participants. For each theme, we briefly outlined: (1) the core problem identified by social media users (e.g., limitations in reporting processes or misuse of intimate content), (2) the associated pain points (e.g., lack of transparency, limited preventative controls, burden on victims), and (3) the improvement directions or feature ideas proposed by users (e.g., clearer triage explanations, warnings for uploaders, stronger controls around image sharing). Following the presentation of each theme, experts were invited to critically reflect on the proposed directions using a structured discussion framework.

For each user proposed solution, experts were asked to consider the following:

- **Effectiveness**: Whether and to what extent the proposed solution addresses the identified problem.
- **Strengths and Risks**: What may work well and what unintended harms or vulnerabilities could arise.
- **Practicality**: Design implications, usability considerations, and technical feasibility.
- **Policy and Ethics**: Regulatory, governance, privacy, and ethical implications.
- **Impact**: How different user groups might experience the solution, including inclusivity and potential unintended consequences.

Data Collection and Analysis

All workshops were conducted with informed consent. Prior to participation, individuals received a detailed Participant Information and Consent Form (PICF) outlining the purpose of the study, the nature of the scenarios, potential risks, and their rights to withdraw at any time. The specific scenario to be discussed in each workshop was shared in advance, and the pre-workshop call provided an opportunity to clarify what participation would involve, discuss expectations, and address any questions before the session.

Workshops were audio-recorded with permission and securely stored. Once transcripts were obtained, recordings were deleted. Outputs from the social media user workshops were analysed using reflexive thematic analysis. The analysis involved iterative reading of transcripts, systematic coding, and ongoing discussion among the research team to refine interpretations and identify patterns across workshops. Analytical memos were used to document emerging insights and decisions throughout the process, supporting transparency and consistency in theme development. Findings were grounded in participants' own words and synthesised into key themes relating to how participants currently interact with and experience social media safety features, the pain points and limitations they encounter when navigating harm, unmet needs that existing platform tools fail to address, and the solutions and improved features participants envisioned to better prevent, detect, and respond to technology-facilitated abuse.

These preliminary themes were then presented to expert participants in subsequent workshops. Experts were invited to critique and reflect on the proposed ideas using a structured framework examining effectiveness, strengths and risks, practicality, policy and ethical considerations, and accessibility and broader impact. This iterative process enabled the refinement of insights by integrating experiential perspectives with feasibility and governance considerations.

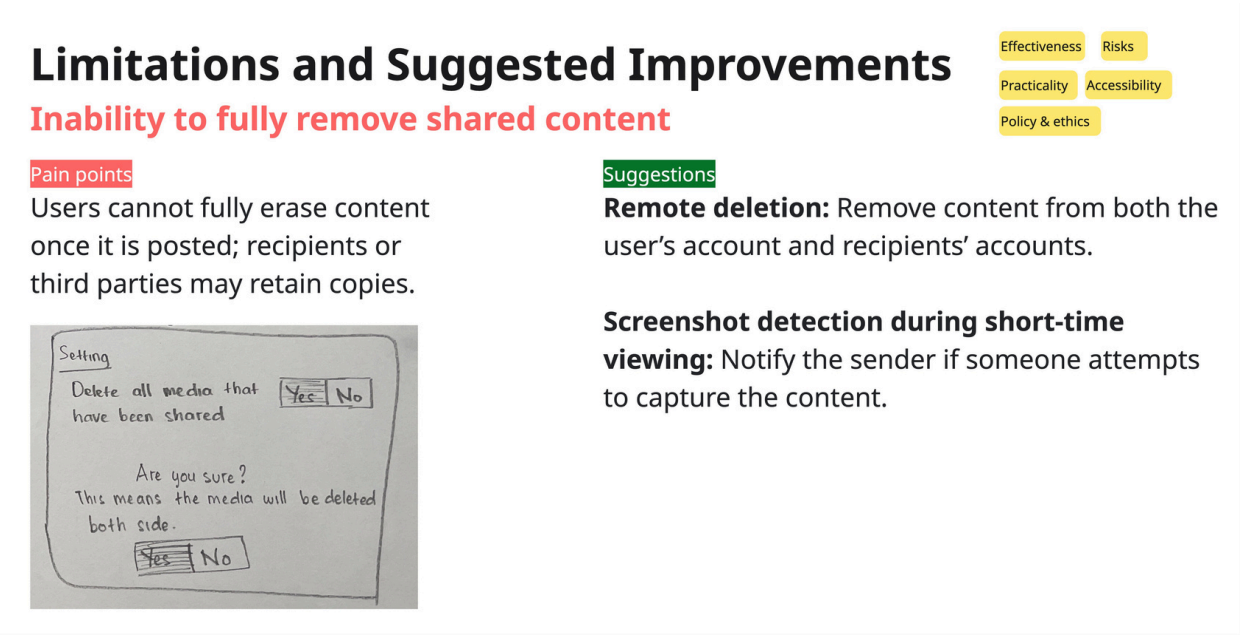


Figure 13: Expert discussion slide example relating to sharing sensitive images.

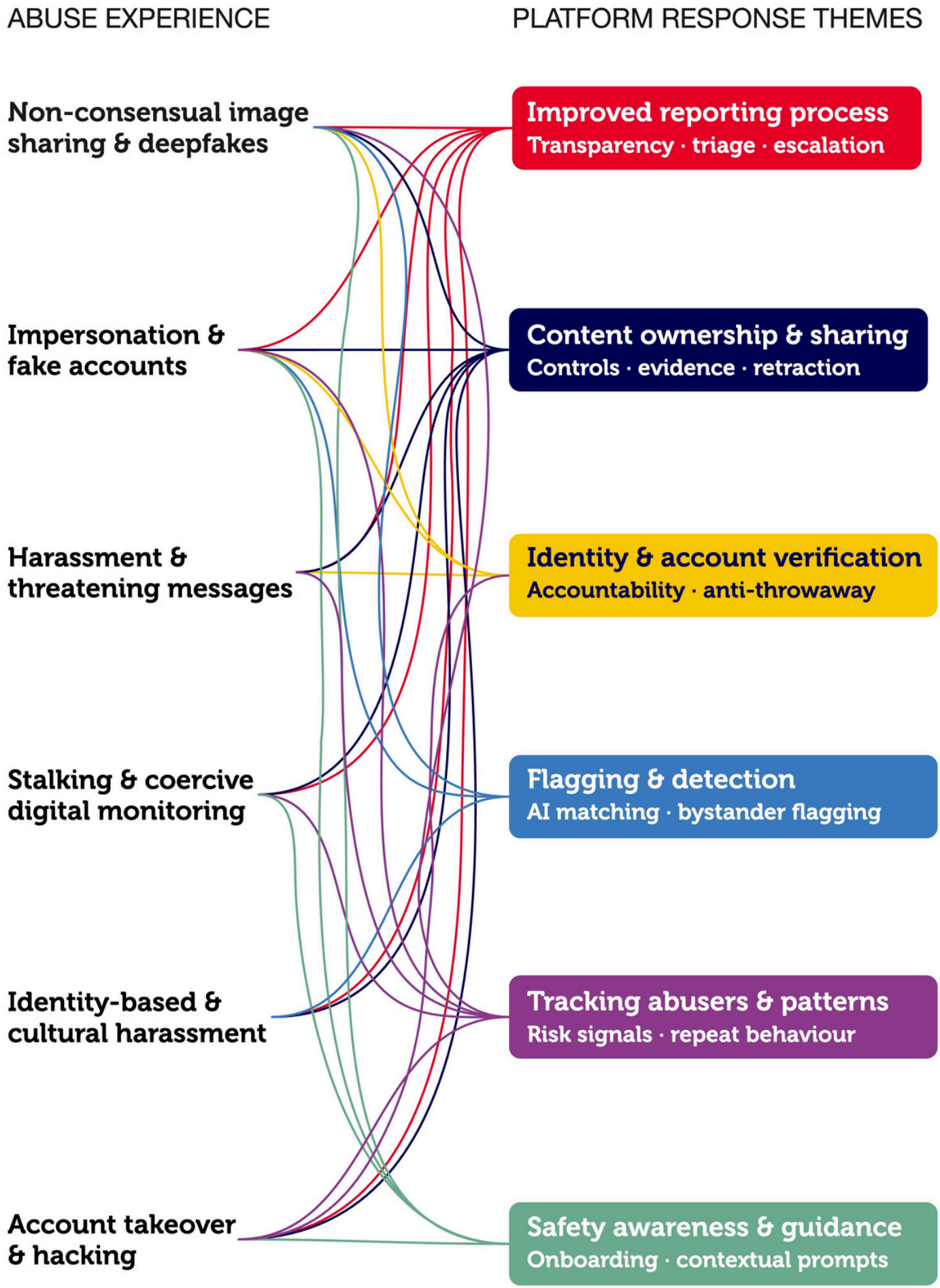


Figure 14: This figure illustrates how platform response themes proposed in the workshops relate to different forms of technology-facilitated abuse. Reporting and content controls were the most widely connected themes, suggesting these are areas where platform improvements may have the broadest potential impact.

Section 4

Findings

Across two rounds of co-design workshops with social media users (n=75) and domain experts (n=21), participants generated a broad set of feature-level and system-level recommendations to strengthen safety on social media platforms for women and gender-diverse people. This section presents the findings organised into six key themes (see Figure 14): Improving the Reporting Process; Content Ownership and Sharing Controls; Identity and Account Verification; Identifying and Flagging Harmful Content; Tracking Abusers and Abusive Behaviours and, Training Users on Safety Features and Awareness.

For each theme, we first describe the proposed features and improvements from the perspective of social media users, illustrated through participant excerpts and user stories. We then outline expert considerations, reflecting on effectiveness, strengths and risks, practicality, policy and ethical implications, and accessibility impacts. Together, these findings provide a structured account of how participants envision safer platform design, and contextualise the design and policy recommendations outlined earlier in this report.

4.1 Improving the Reporting Process

Discussions about reporting were particularly prominent in scenarios involving non-consensual intimate image sharing (Scenario G), persistent unwanted contact (Scenarios A and C), impersonation and account misuse (Scenarios B and F), stalking and threats (Scenario L), and cross-platform harassment by an ex-partner (Scenario D) (see full scenario descriptions in Scenarios). In these contexts, reporting was often described as the primary formal action available to users.

Figure 15: Key themes and ideas proposed by participants for improving the social media reporting process.

However, participants consistently characterised existing reporting mechanisms as slow, opaque, and emotionally discouraging. Many described submitting reports and then "hearing nothing back", feeling unsure whether their concerns were reviewed, dismissed, or simply lost within automated systems, particularly in high-risk situations such as repeated re-uploads of intimate content or escalating threats. This uncertainty intensified distress and reduced trust in the platform. Some participants described reporting as burdensome, particularly when required to repeatedly document harm across multiple accounts or platforms.

Rather than requesting entirely new safety tools, participants focused on improving the reporting experience itself. They wanted clearer categories that matched their situations, simpler steps during moments of distress, visible progress updates, and reassurance that a real person or accountable system was taking their case seriously. Across workshops, the emphasis shifted from simply enabling reports to making the process feel responsive, transparent, and supportive. The following features reflect these needs.

- Clearer reporting categories and contextual descriptions (widespread request)
- Human-AI moderation support (widespread request)
- Tracking and progress visibility after reporting (widespread request)
- Transparent actions and outcomes (widespread request)
- Delegation or trusted-person support (occasional request)
- Access to timely assistance or escalation channels (common request)

*Widespread: almost everyone said this, common: a lot of people said this, occasional: some mentioned, isolated: one-off idea.

4.1.1 Clearer reporting categories and contextual descriptions

Participants frequently described difficulty matching their experiences to the limited reporting categories provided by platforms. Harassment, impersonation, or identity-based abuse often fell into ambiguous or overly broad options, which made reports feel misclassified or dismissed. Several participants noted that subtle or contextual harms, such as coded language, cultural mockery, or implied threats, were especially hard to capture through predefined checkboxes. Without space to explain the situation in their own words, users worried that automated systems or reviewers would misunderstand the severity of the issue, resulting in inaction.

To address this, participants proposed more specific and representative categories, along with opportunities to add short written explanations, examples, or supporting context. Some suggested layered flows with main categories followed by subcategories to better narrow the problem, while others asked for open text fields to describe nuance that AI might not detect. These additions were seen as ways to improve both the accuracy of moderation decisions and users' confidence that their concerns were properly understood. One participant shared how some platforms can extend reporting categories with a space for explanation:

"It would help if there was an 'other' field where you could write details. Maybe AI could sort those words into categories, and then a human could review the ones that come up often..."

23-year-old Asian woman, active social media user who creates and shares content, reshares content, and consumes content

*Participant avatars are AI-generated illustrations and do not represent the appearance or identity of the individuals quoted.

EXPERT CONSIDERATIONS

Experts suggested that clearer categories combined with lightweight contextual input and appropriate human review may improve both accuracy and user trust, while remaining feasible for large-scale moderation.

Effectiveness Experts agreed that poorly defined or overly broad categories can reduce moderation accuracy, as reports may be routed to the wrong queues or misinterpreted by automated systems, leading to inconsistent or delayed outcomes.

Practicality While adding more categories or free-text fields can improve precision, experts cautioned that this also increases review complexity and workload. They suggested the need for better triage mechanisms, such as AI-assisted sorting, to ensure reports remain manageable at scale.

Impact Experts highlighted that many harms, particularly identity-based, cultural, or contextual abuse, are difficult for automated systems alone to interpret. Allowing users to describe situations in their own words and enabling human oversight can help ensure these experiences are more fairly assessed.

4.1.2 Human-AI moderation support

Although reporting tools were available, many participants felt that their concerns disappeared into automated systems with little reassurance that anyone had actually reviewed their case. Responses were often described as generic, delayed, or impersonal, which reduced trust in the platform and discouraged further reporting. Several participants expressed frustration at 'chatbot loops' or automated replies that did not fully understand nuanced or identity-based harms.

In response, participants repeatedly called for ways to reach a real person, particularly for serious or sensitive situations such as harassment, impersonation, or intimate image abuse. Rather than relying solely on automated moderation, they wanted clearer escalation pathways where complex cases could be reviewed by trained human moderators. Some also recognised that AI could still play a role in triaging or routing reports quickly, but emphasised that human judgement and empathy are critical for fair decisions and emotional reassurance.

Participants often described AI not as something to replace people, but as a way to quickly route their issue to the right support:

"Maybe the AI could first just ask what the problem is and then redirect you to whoever is in charge. Like, if I say someone's impersonating me, it recognises that category and transfers me to an actual human who knows what to do. So we're kind of cutting out the middleman and not waiting around for approval."

21-year-old Asian woman, social media user who reshares and consumes others' content

Others emphasised that fully automated systems struggle with nuance, especially in more serious or contextual cases:

> *"If it's really serious, a human on the other end could review it instead of it being purely computerised, because sometimes it's hard to get it right with just what a computer can catch."*
>
> 23-year-old Asian woman, active social media user who creates and shares content, reshares content, and consumes content

EXPERT CONSIDERATIONS

Overall, experts viewed human-AI collaboration as a practical direction, with AI helping manage scale and humans providing judgement and care in more complex situations, provided appropriate training and privacy safeguards are in place.

Effectiveness

Experts agreed that AI systems alone often struggle to interpret contextual, cultural, or identity-based harms. Human moderators are better positioned to assess nuance and make proportionate decisions in complex cases.

Practicality

Fully human moderation for all reports is not feasible at scale. Experts agreed on the idea of hybrid approaches where AI supports triage and prioritisation, with human reviewers focused on high-risk or sensitive cases.

Policy and ethics

Enabling direct human contact may encourage users to share sensitive or traumatic experiences, which raises concerns about whether moderators are adequately trained to respond safely and how such personal information is handled and protected.

4.1.3 Tracking and progress visibility after reporting

A recurring frustration across workshops was the lack of visibility once a report was submitted. Participants described reporting harmful posts or accounts and then hearing nothing back, sometimes for days or weeks. Without updates, it was unclear whether the report had been received, reviewed, or acted upon. Several participants also noted that they did not know whether a dashboard or dedicated space even existed to check on the status of their reports. This uncertainty left people feeling ignored, unsupported, or forced to repeatedly report the same issue or ask friends to help escalate it.

Participants therefore called for clearer feedback and progress tracking throughout the process. Suggestions included acknowledgement messages, estimated timelines, status dashboards, and notifications explaining what actions were taken. Rather than a one-off submission, reporting was envisioned as an ongoing, transparent process that shows where a case sits and what happens next.

Participants described wanting clearer, tangible feedback that shows where their report sits and what happens next, ranging from simple timelines to more detailed dashboards:

"There's a report dashboard [persona of the scenario] can check with the details of her report and its status, whether it's active or under review. It shows what's going on, the possible timeline for how it will be processed, and then updates after a few days about what action was taken. Knowing that the report is heard and something has happened makes her feel secure and relieved, and able to use the platform safely again."

21-year-old Asian woman, social media user who creates and posts content, consumes others' content and moderates social media accounts or groups

EXPERT CONSIDERATIONS

Experts viewed greater transparency and feedback as a practical, low-risk improvement that could strengthen trust in the reporting process.

Effectiveness Clear acknowledgements and status updates may reduce anxiety and repeated reporting by helping users understand that action is underway.

Practicality Implementing timelines, dashboards, or automated updates are viewed as technically feasible and compatible with existing ticketing or case-management systems, and could also support victim-survivors in holding platforms to account.

Impact Transparent progress indicators may particularly benefit users under stress, by lowering cognitive load and avoiding the need to navigate complex help channels or chase responses.

4.1.4 Transparent actions and outcomes

Beyond simply tracking the status of a report, many participants wanted clearer insight into what happens behind the scenes once a report is submitted. While some platforms provided generic acknowledgements such as 'thanks for your report', participants described feeling left in the dark about how decisions were made, who reviewed the content, or why certain actions were or were not taken. This lack of transparency often led to uncertainty and mistrust, particularly when harmful content remained visible or when reports were dismissed without explanation.

Participants suggested that platforms communicate not only progress, but also the reasoning behind outcomes. They wanted to understand how reports were triaged, whether AI or human moderators were involved, what evidence was considered, and what criteria determined removal, warnings, or inaction. Making these processes visible was seen as a way to feel heard, regain a sense of control, and hold platforms accountable, rather than feeling that reports disappeared into an invisible system.

Participants emphasised that beyond simple status updates, they wanted clearer insight into who was handling their case and how decisions were being made behind the scenes.

"Sometimes you don't have a clear understanding of where the case is at, whether there's a case officer assigned, or what's happening in the background. It makes me really anxious if I don't know what's going on, and it doesn't help if I want to take further action. I would want to see what's happening."

35-year-old Asian woman, active social media user who creates and posts content, reshares content, participates in discussions and moderates social media accounts

EXPERT CONSIDERATIONS

Experts agreed that clearer explanations of decisions and outcomes could improve trust and reduce frustration, particularly when users currently feel unsure about what happens after they submit a report.

Effectiveness Clear explanations of how decisions are made (for example, why content was or was not removed) may improve user trust and reduce repeated or duplicate reports driven by uncertainty.

Practicality Providing meaningful explanations for each case may require structured moderation logs or templated decision summaries, balancing clarity with the operational burden on moderation teams.

Impact Providing updates and decision rationales can reduce anxiety and emotional burden, particularly for people already under stress, helping them feel heard and supported rather than ignored by an opaque system.

4.1.5 Delegation or trusted-person support

Several participants described how managing the aftermath of reporting such as gathering evidence, following up on updates, and navigating platform processes can feel overwhelming, especially when someone is already stressed or affected by harassment or experiencing trauma. Friends or family frequently step in informally to help, but there are few formal ways to share responsibility or coordinate support within the platform. As a result, participants expressed interest in features that would allow trusted contacts to assist with managing reports, tracking progress, or handling communications, reducing both practical and emotional burden.

"I envision the feature to give you the freedom to pass on the whole process to someone you trust, like a friend or family member. Kind of like a power of attorney. To take the burden off, because you're already feeling distressed."

29-year-old European woman, active social media user who creates and posts content, consumes content and moderates social media accounts

EXPERT CONSIDERATIONS

Experts framed delegation as enabling supportive accompaniment through the reporting journey.

Effectiveness Clear explanations of how decisions are made (for example, why content was or was not removed) may improve user trust and reduce repeated or duplicate reports driven by uncertainty.

Strengths and risks Delegation must avoid exposing sensitive details to unsafe individuals or escalating harm if abusers (especially in abuse inflicted by current/ex-partners) gain visibility into the reporting process.

Impact Providing options to share the burden including choosing what information is shared, was seen as a way to meet users with empathy during crisis moments.

4.1.6 Access to timely assistance or escalation channels

Beyond reporting and tracking, participants emphasised the need for rapid intervention pathways, particularly in cases involving private or intimate image abuse. The speed at which such content spreads heighten anxiety and distress, with participants expressing that delayed platform responses could significantly amplify harm.

For some participants, timely acknowledgement alone was insufficient; they wanted immediate containment measures to minimise further exposure. They discussed the need for escalation beyond the platform when harm crosses into criminal or legal domains. In particular, some expected direct pathways to law enforcement where appropriate:

"I was hoping that if there was a feature where I could report this account, and that report would go directly to the police so they could track this person. I think I might feel safe sooner"

27-year-old Asian woman, active social media user who creates and posts content, and reshares others' content

EXPERT CONSIDERATIONS

Experts supported timely escalation pathways for high-risk cases but emphasised the need for clear severity triage and structured human oversight. They cautioned that without careful scoping and coordination with external authorities, such mechanisms could create unrealistic expectations or operational strain.

Effectiveness Escalation pathways must be clearly tiered according to severity. Cases involving credible threats, repeated harassment, or non-consensual intimate image distribution require faster triage and prioritised review.

Strengths and risks

Providing direct access to human moderators or external authorities can improve user reassurance and support documentation for potential legal action. At the same time, experts noted risks including reports being escalated to law enforcement prematurely or without sufficient evidence, the potential for malicious misuse of escalation pathways, and users developing expectations that platforms can guarantee an immediate police response. Clear criteria and safeguards would therefore be necessary to prevent misuse and ensure escalation pathways function as intended.

Practicality

Experts raised concerns about the operational burden of real-time human support or 24/7 hotlines. They suggested that scalable escalation may require AI-supported triage combined with human review for high-risk cases, as well as careful coordination with external bodies across jurisdictions.

Policy and ethics

Experts highlighted the need for clear data-sharing protocols when cases are escalated beyond the platform. Escalation involving police or regulators must comply with privacy laws and ensure informed user consent, particularly in sensitive cases such as image-based abuse.

Impact

Users experiencing distress may require both enforcement pathways and wellbeing support. Escalation features should therefore be clearly signposted, easy to understand, and accompanied by links to appropriate counselling or support services to reduce anxiety and improve perceived safety.

4.2 Content Ownership and Sharing Controls

Participants described that harm often escalates not only because harmful content exists, but because it can be quickly copied, saved, and redistributed in ways that are hard to trace or reverse. Once a photo, message, or post is screenshotted, downloaded, or forwarded, it can move outside the platform's control and remain accessible indefinitely. This made people feel that current safety options are limited because they mainly react after harm has spread, rather than preventing misuse at the point of sharing.

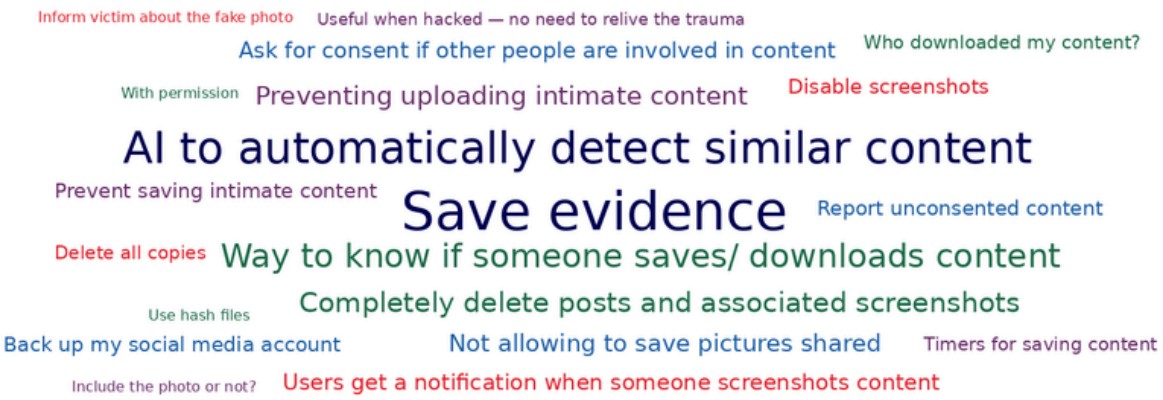


Figure 16: Key themes and ideas proposed by participants around controlling, preserving, and managing content on social media platforms

These concerns were reflected across multiple scenarios, including non-consensual intimate image sharing (Scenarios G, C, and D), AI-generated fabricated images (Scenario I), unwanted or escalating private messaging involving explicit or threatening content (Scenarios A, C, and H), and impersonation through hacked or fake accounts (Scenarios D and F) (see full scenario descriptions in Scenarios). In each case, the speed and ease with which content could be captured, edited, or re-uploaded intensified the harm and reduced users' sense of control.

In response, participants described a need for practical safeguards built into everyday content interactions. They wanted clearer control over who can save content, whether saving should be allowed at all, and how to limit redistribution of sensitive images that are shared consensually. They also highlighted tensions in what 'safety' means in practice, for example wanting to prevent misuse through screenshot restrictions while also needing reliable ways to preserve evidence if the situation escalates. Features users asked for includes the following:

- Controls to restrict screenshotting, screen recording, downloading, or saving content, or to notify users when this happens. (widespread request)
- Options to preserve or export evidence in a reliable way, including timestamps and platform-verified records. (common request)
- Ability to delete or retract sent content from both ends, especially intimate images or sensitive messages. (common request)
- Restrictions or friction for sharing intimate or explicit content, such as warning screens, blur-by-default, or timers. (occasional request)
- Consent prompts or checks when uploading or posting content that involves another person. (occasional request)
- Personal backup or archival tools so people can leave, restart, or recover accounts without losing important memories or content. (occasional request)

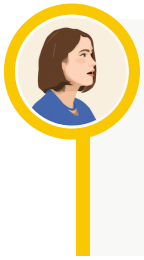
*Widespread: almost everyone said this, common: a lot of people said this, occasional: some mentioned, isolated: one-off idea.

4.2.1 Restrictions and permissions for saving, downloading, or screenshotting content

Participants repeatedly described that once content is saved by someone else, it becomes difficult to contain harm. Screenshots and downloads were seen as a key point where private images, close-friends posts, or sensitive messages can be redistributed without consent. People also described uncertainty about who has accessed their content, and fear that saved copies can be used later for threats, coercion, or ongoing harassment. Several participants pointed out that even if blocking or restricting helps in the moment, it does not prevent someone from keeping a permanent copy once they have downloaded or captured it.

Participants suggested stronger controls at the point where content can be saved. This included restricting screenshots or downloads entirely for certain posts, requiring permission before saving media, and notifying the original user when someone attempts to screenshot or screen record. Some participants also proposed 'view-only' modes, timed viewing, or blurring that reduces the chance of capturing content quickly. A few participants recognised limitations, noting that a determined person could still use another device to photograph a screen, but they still saw platform-level restrictions as a meaningful barrier that reduces casual misuse and makes suspicious behaviour more visible.

As one participant suggested:



"I want platforms to disable screenshots for 'sharing controlled' posts so that users can further protect their privacy."

29-year-old European woman, active social media user who creates, shares and consumes content

EXPERT CONSIDERATIONS

Overall, experts viewed these controls as useful safeguards that may reduce casual misuse and increase awareness, but not as complete protections, suggesting they be considered alongside evidence preservation and broader moderation measures.

Effectiveness

Measures such as screenshot notifications, download restrictions, or automatic detection of reused content could increase friction and awareness, helping users notice when their content is being copied or misused. However, they cautioned that these features are unlikely to fully prevent misuse, as determined actors can still capture content through alternative means (for example, external devices or third-party tools). As such, these mechanisms may reduce casual harm but not eliminate deliberate abuse.

Strengths and risks

Providing notifications, permissions, or automated logging could give users greater visibility and a sense of control over how their content circulates, and may support earlier intervention or documentation. At the same time, experts noted a tension between restricting capture and enabling victims to collect evidence. Preventing screenshots entirely could unintentionally limit a person's ability to preserve proof of harassment or abuse.

Practicality

From a technical standpoint, experts highlighted that many controls operate only at the app level and are difficult to enforce reliably at the device level. Once content appears on a screen, it is challenging to stop copying altogether. They suggested that more feasible approaches may include configurable permissions (e.g., per chat or per post settings) or notifications rather than hard blocks. At the time of this report, some platforms have begun experimenting with features that notify users when screenshots are taken or restrict what can be captured (e.g., displaying blank screens), indicating emerging directions for adding layers of privacy protection, although such measures remain limited in scope and effectiveness.

Impact

Experts also reflected on the potential burden placed on users. Frequent warnings, confirmations, or notifications may become normalised and ignored over time, reducing their usefulness or adding cognitive load. Overly complex controls could disproportionately affect people with lower digital literacy or those already managing stressful situations.

4.2.2 Options to preserve or export evidence in a reliable way

While participants wanted to limit how others could copy or redistribute their content, many also described the opposite need when they were facing harassment or abuse: the ability to reliably preserve proof. Messages, images, and posts could be quickly deleted by perpetrators, accounts could disappear, or content could be altered, leaving people without a record of what had happened. Several participants described feeling rushed to manually take screenshots before material vanished, which was stressful and easy to miss.

Participants also expressed uncertainty about whether screenshots alone would be considered trustworthy or sufficient, particularly if they needed to escalate a case to the platform, the Australian eSafety Commissioner, or police. As a result, current approaches felt fragile, time-consuming, and dependent on users remembering to document incidents themselves.

To address this, participants suggested built-in evidence tools that automatically or reliably capture key information when harmful behaviour occurs. Ideas included secure logs of incidents, account details, and platform-verified records that could be compiled into a single document or export. Some described features that automatically save reports or generate an authenticated summary of the harmful incident, including what occurred, who was involved, and when, that could be shared with authorities if needed. These were framed as protective supports that reduce the burden on individuals and make it easier to act when situations escalate.

As one participant explained:

"The first step I would take, without a second thought, is screenshotting and saving the imagery and offensive comments... because just as quickly as someone can post or send them, they can delete them. It's important to have that evidence... as the first step in making a police report or filing a complaint."

29-year-old European woman, active social media user who creates, shares and consumes content

EXPERT CONSIDERATIONS

Overall, experts viewed evidence tools as supportive safeguards that may strengthen users' ability to seek help and escalate harm, particularly when designed to be simple, secure, and optional rather than burdensome.

Effectiveness

Platform-generated logs, timestamps, and verified records were seen as potentially more credible and complete than manual screenshots, and may reduce the burden on users to collect evidence during stressful situations.

Strengths and risks

Centralising evidence can support reporting and investigations, but may also create expectations that platforms store and manage sensitive personal information for longer periods, which could introduce additional responsibility and risk.

Practicality Automating capture of key metadata (for example, time, account ID, message history) was considered technically feasible, though experts noted the need for clear boundaries around what is stored and how long records are retained.

Policy and ethics Centralising evidence introduces responsibilities around secure storage, access control, and retention.

Impact Providing simple, one-step exports or summaries may reduce cognitive load and make it easier for users in distress to seek help without needing complex technical knowledge.

4.2.3 Ability to delete or retract sent content from both ends

Although many platforms already allow messages to be deleted or unsent, participants described these options as limited and unreliable in situations involving harassment, breakups, hacking, or coercion. Existing features often only remove content from the sender's view or work within short time windows, leaving copies accessible to the recipient. Once sensitive images or messages had been shared, participants felt they had little control over what persisted, especially if the other person had already saved or redistributed the material.

Participants also described the emotional burden of content remaining visible in their history. Having to manually scroll through and delete messages one by one was described as time-consuming and distressing, particularly when trying to distance themselves from harmful experiences.

Participants asked for stronger forms of retraction that could support complete removal of content and reduce the likelihood of future misuse. Suggestions included deleting messages or images from both sides of a conversation, bulk or time-range deletion such as removing all content from a specific period of engagement with the abuser, and options to automatically remove sensitive media that had already been shared. These features were framed as ways to regain control, prevent continued threats or redistribution, and reduce the emotional impact of revisiting harmful material.

One participant expressed their need:

"I want to delete all history [of our chat] on both sides, so [persona of the scenario] doesn't have evidence that I sent it [the sexually explicit material]."

29-year-old Asian woman, social media user who mostly consumes content

EXPERT CONSIDERATIONS

Overall, experts viewed stronger retraction tools as supportive mechanisms that may help users regain control and reduce harm, while noting that their effectiveness depends on clear expectations about what can and cannot be fully removed.

Effectiveness	Removing content from both sides may reduce ongoing exposure and limit immediate reuse, particularly in cases where harm is contained within the platform.
Strengths and risks	Retraction tools can give users a sense of control and closure, but may create false confidence if recipients have already saved copies or captured content externally.
Practicality	Experts noted that deletion can usually be enforced within platform storage, but becomes harder once files are downloaded, screenshotted, or shared elsewhere.
Policy and ethics	Allowing retroactive deletion must balance user protection with transparency and record-keeping needs, particularly where content may be relevant to investigations or disputes.
Impact	Bulk or time-based deletion may reduce cognitive and emotional burden, especially for users managing distressing situations or large volumes of content.

4.2.4 Restrictions or friction for sharing intimate or explicit content

Participants described that many harmful situations begin at the moment sensitive images or messages are shared, often impulsively, under pressure, or without fully considering how the content might later be saved, forwarded, or used against them. Once intimate or explicit material is shared, control quickly diminishes, particularly if trust breaks down or relationships change. Several participants reflected that existing platforms make sharing quick and effortless, but offer few safeguards that encourage people to pause or reconsider before sending something sensitive.

This ease of sharing was seen as contributing to regret and vulnerability. Participants noted that even if later protections such as blocking or deletion exist, they often come too late once content has already circulated.

To reduce these risks, participants suggested adding preventative friction during the sharing process itself. Ideas included warning screens before sending intimate media, blur-by-default previews for the recipient, timers or one-time viewing options, and prompts that ask for confirmation or consent before content is posted or forwarded. Others suggested requiring permission before downloading or resharing sensitive images, or automatically limiting how widely such content can be distributed. These measures were framed as small interruptions that create space to reflect, rather than hard restrictions that block communication entirely.

One participant suggested to ask permission for downloads:

"If someone wants to download a picture I shared, they should ask for my permission first... not to stop sharing completely, but to help me think twice about whether it's worth sending."

29-year-old Asian woman, social media user who mostly consumes content

EXPERT CONSIDERATIONS

Overall, experts suggested that introducing friction can support safer decisions, while recognising that technical and behavioural limits mean these measures should complement, rather than replace, other safeguards.

Effectiveness	Warning screens, timers, and confirmation prompts may discourage impulsive sharing and reduce accidental or casual redistribution, particularly for sensitive content.
Strengths and risks	Lightweight friction can promote safer behaviour without removing user choice. However, implementing warnings or blur-by-default features depends on automated detection that may be imperfect, and frequent prompts risk being ignored or adding friction to everyday communication.
Practicality	These measures are generally feasible at the interface level and can be applied selectively, for instance, to certain media types or settings.
Policy and ethics	Providing warnings or automated checks may require platforms to analyse private content, raising questions about user consent, surveillance, and how sensitive material is assessed.
Impact	Simple, consistent controls may support safer decision-making, but overly complex settings or repeated interruptions could add cognitive load, particularly for users already under stress.

4.2.5 Consent prompts or checks when uploading or posting content that involves another person

A smaller group of participants focused on situations where harm occurred because content that involved them was posted by someone else without their knowledge or permission. This included screenshots of private conversations, shared photos, or images that could be sensitive, intimate, or personally identifiable. Once uploaded, participants often only discovered the content after it had already circulated, leaving them to react rather than prevent the situation.

Participants described feeling that current platforms place the most responsibility on the person affected to report or remove content after the fact, rather than creating safeguards before something is posted. This reactive approach was considered stressful and insufficient, particularly when the material was sensitive or potentially harmful.

To address this, some participants suggested introducing consent checks at the point of upload when content clearly involves another identifiable person. Ideas included prompts asking whether permission had been obtained, notifications to the person depicted, or temporary holds that allow them to review or approve certain posts. These were framed as preventative measures that create a pause before sharing and reinforce expectations of consent, rather than blocking posting entirely.

One participant suggested to prompt the uploader about content involving other people:

"Maybe there can be features when you're about to post something that involves other people... like a question or some kind of preventative measure. If it's private or intimate, maybe it asks whether you've got the other person's consent before you can post it."

35-year-old Asian woman, active social media user who creates and shares content, engages in discussions, and comments on posts

Another participant suggested to warn the uploader if a similar image is reported in the platform before:

"If it's similar to something that's been flagged before, it could send a pop-up saying this looks like previously reported content and needs review before uploading... so it goes through moderation first. If it was flagged by mistake, they can still upload it."

23-year-old Asian woman, active social media user who creates and shares content, consumes content, engages in discussions, and comments on posts

EXPERT CONSIDERATIONS

Experts viewed consent checks as supportive preventative cues that may encourage more respectful sharing practices, while recognising that they are unlikely to prevent intentional misuse and are best considered as complementary safeguards.

Effectiveness Consent prompts may reduce accidental or thoughtless sharing and encourage more respectful behaviour, particularly in close relationships or peer contexts.

Strengths and risks These checks reinforce norms around permission and accountability, but may be less effective in situations involving deliberate abuse or malicious intent.

Practicality Automatically identifying when a post involves another person can be technically challenging and may rely on tagging, facial recognition, or user input, which may be inconsistent or inaccurate.

Policy and ethics Detecting or flagging content that features specific individuals may require analysing images or identities, raising questions about consent, biometric processing, and how such detection is governed.

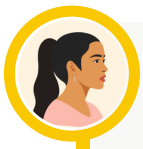
Impact Additional prompts may add friction to everyday posting and could become burdensome if triggered too frequently, suggesting the need for selective or configurable application.

4.2.6 Personal backup or archival tools for safer exit and recovery

Few participants discussed situations where leaving or closing an account felt necessary for safety or wellbeing, such as after harassment, hacking, or ongoing abuse. However, doing so often meant losing years of photos, messages, connections, and personal history. This created a difficult trade-off: staying on a platform that felt unsafe, or leaving and losing important memories and content.

Participants described feeling attached to their accounts not only socially but personally, as records of relationships, milestones, and creative work. As a result, starting over or deleting an account could feel overwhelming and emotionally costly, even when it might otherwise be the safest option. Some participants also raised the need for bulk deletion tools that would allow them to remove large volumes of content, such as all posts or messages from a specific period, in a single action, rather than having to manually revisit and delete harmful or distressing material one item at a time.

To reduce this barrier, participants suggested clearer and more flexible backup or archival options that allow people to safely export, store, or transfer their content before leaving. Ideas included downloading posts and media in bulk, saving messages or timelines, or transferring core information to a new account. These tools were framed as ways to give people more freedom to exit harmful situations without feeling that they must sacrifice their digital history.



"I want [platform] to enable a feature that lets you delete posts from a certain time period in one go... so I don't have to relive traumatic events or messages if there are hundreds"

32-year-old Asian woman, active social media user who creates and shares content, and consumes content

EXPERT CONSIDERATIONS

Experts viewed archival and backup tools as supportive recovery features that may help users leave or rebuild safely, reinforcing a sense of control without directly interfering with everyday use.

Effectiveness	Backup or archival tools may lower the barrier to leaving unsafe situations by allowing users to retain important content and memories.
Strengths and risks	These features support user autonomy and recovery, though exporting large amounts of personal data may introduce risks if files are stored insecurely or shared unintentionally.
Practicality	Bulk export and transfer functions are technically feasible and already supported in some form on many platforms.
Policy and ethics	Providing data portability aligns with broader expectations around user ownership and control of personal information.

4.3

Identity and Account Verification

Across nearly all workshops, identity verification emerged as a core concern linked to accountability and repeat abuse. Participants repeatedly described how easily fake, duplicate, or impersonation accounts could be created, allowing perpetrators to bypass account blocks, return after bans, or misuse someone else's identity. As a result, identity verification was discussed not as an additional feature, but as a foundational safeguard that could increase accountability, deter throwaway accounts, and provide stronger recourse when harm occurs. The following subsections outline the specific verification-related approaches raised by users and examined with experts. Features users asked for:

- Identity or ID-based verification during account creation (widespread request)
- Limits on multiple or throwaway accounts (occasional request)
- Verification for impersonation cases (occasional request)

*Widespread: almost everyone said this, common: a lot of people said this, occasional: some mentioned, isolated: one-off idea.

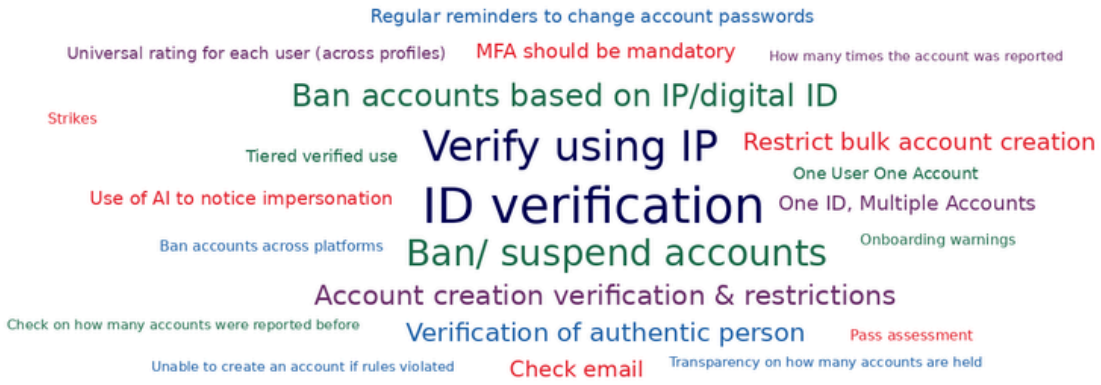


Figure 17: Key themes and ideas proposed by participants for strengthening account authenticity and reducing opportunities for repeat or anonymous abuse.

4.3.1 Identity or ID-based verification during account creation

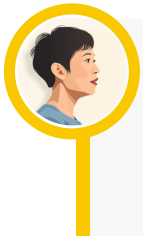
Across many workshops, participants described how easily abusive or fake accounts could be created. Blocking or reporting often felt ineffective because the same person could quickly return under a new profile, sometimes within minutes. This created a cycle where harmful behaviour resumed despite users taking protective actions, leaving participants feeling that current safety tools were temporary and reactive rather than preventative.

Participants also pointed to the broader issue of anonymity and limited accountability. Fake emails, fabricated names, and disposable accounts made it difficult to trace who was behind harmful behaviour or to demonstrate that an account was impersonating someone. Without

clearer ways to link profiles to a real person, users felt there were few meaningful consequences for repeat offenders.

To address these frustrations, many participants suggested linking accounts to some form of verifiable identity, such as a government or digital ID, during account creation. Verification was framed as a basic gatekeeping step that could discourage anonymous misuse, increase accountability, and reduce the ease of creating throwaway or fraudulent profiles.

Participants described verification as a straightforward safeguard:



"...every time you generate an account then you need to be ID verified. I think it will be able to hold more people accountable for what they say and do..."

32-year-old Asian woman, active creator, sharer, and moderator on social media

EXPERT CONSIDERATIONS

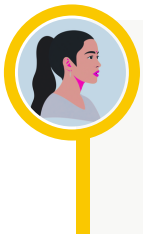
Experts acknowledged that some platforms already use forms of identity verification and agreed it can support accountability. However, they cautioned against making verification mandatory or universal, citing privacy, data security, and exclusion risks. They suggested offering verification as an opt-in safeguard and combining it with stronger enforcement of repeat offenders and improved moderation tools.

Effectiveness	May deter casual or large-scale account creation by adding friction compared to the near-instant sign-up processes common on many platforms. However, it is unlikely to stop determined offenders, who may use alternative credentials, shared accounts, or other workarounds to continue their behaviour.
Strengths and risks	Could improve accountability, but introduces significant privacy and data security risks, particularly around how sensitive identity information is collected, stored, and managed by platforms.
Practicality	Complex to implement consistently across jurisdictions. Countries vary widely in the types of identity documents available, and not all users have access to government-issued identification.
Policy and ethics	Platforms must consider the privacy implications of collecting and storing identity information, including risks of surveillance and centralised identity tracking, and should be transparent with users about how such data is managed and protected.
Impact	May exclude people without formal ID or those who rely on anonymity for safety.

4.3.2 Limits on multiple or throwaway accounts

In some workshops, participants described how easy it was to set up new or duplicate accounts using different emails or details, allowing the same person to repeatedly return under different profiles. This made accounts feel disposable and undermined efforts to stop harmful behaviour, as issues resurfaced through fresh accounts rather than being contained at the source.

To address this, many participants suggested restricting how many accounts a person can create or linking related accounts together. Proposals included tying accounts to a device, mobile number, or form of ID, detecting repeated accounts from the same source, or making the number of accounts more visible so patterns of misuse are easier to spot.



"Each profile could show how many accounts a person has... maybe limit it to five, so one person can't keep creating more."

30-year-old Asian woman, active social media participant who shares and comments

EXPERT CONSIDERATIONS

Experts noted that account limits can add friction but are unlikely to prevent repeat abuse on their own. They discussed pairing limits with additional ways to recognise related accounts and support consistent moderation, reinforcing similar ideas raised by users.

Strengths and risks These features can add accountability while still allowing flexibility for users who legitimately manage more than one account.

Practicality Implementation may require reliable ways to link accounts (e.g., email, phone number, or verified ID), which may be imperfect or easy to circumvent.

Impact Strict limits may affect people who share devices or need separate accounts for work, family, or safety reasons.

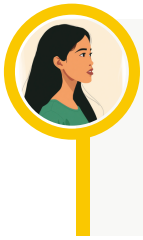
4.3.3 Verification for impersonation cases

Some participants expressed the stress and uncertainty of being impersonated online. When someone created a look-alike profile using their name or photos, the original person often struggled to prove impersonation. Participants said this made reporting impersonating accounts feel slow and frustrating, as platforms appeared to rely on profile photos or basic information that could easily be copied. As a result, the burden of proof often fell on the victim rather than the impersonator.

To address this, participants suggested stronger ways for individuals to verify their identity specifically during impersonation disputes. Rather than relying only on static images or form

submissions, they proposed live selfie checks, short video verification, or time-bound photo matching that could demonstrate the presence of the real person. These approaches were seen as ways to quickly confirm authenticity for the legitimate user while creating additional barriers for impersonators or hackers, as live or time-sensitive verification steps are considerably harder to replicate fraudulently. Participants also noted that with the growing availability of AI-generated images and deepfakes, simple photo uploads may no longer be sufficient proof of identity.

As one participant suggested:



"Maybe take the photo right then and there using the actual camera, rather than uploading an existing one, because otherwise they can fake it again."

29-year-old Asian woman, active social media user who creates and shares content

EXPERT CONSIDERATIONS

Experts described enhanced verification in impersonation cases as a targeted safeguard rather than a universal requirement. When used selectively during disputes, such measures may help legitimate users re-establish control over their identity and reduce opportunities for impersonation, provided strong privacy protections and accessible alternatives are in place.

Effectiveness	Could help establish which account belongs to the legitimate person and support faster resolution of impersonation reports. Live or time-bound checks may also discourage casual or low-effort impersonation attempts, where perpetrators rely on the ease of copying a profile photo or replicating basic account details.
Strengths and risks	May restore a sense of control for victims, but introduces privacy and data security risks, particularly around how biometric data collected during live or time-bound verification, such as facial images or video captures, is stored, managed, and protected by platforms
Practicality	Current platform systems focus mainly on reporting workflows rather than real-time verification. On Meta platforms, users report impersonation through in-app forms and may be asked to provide identifying information such as a government ID during follow-up steps [22, 23]. On TikTok [31], users select 'Pretending to Be Someone' and may upload identification documents as part of the reporting process. These approaches rely on static documentation rather than live confirmation, which experts felt limits their effectiveness against sophisticated impersonation.
Policy and ethics	Requires explicit consent, minimal data collection, and clear policies on retention and deletion to avoid over-collection or misuse of personal information.
Impact	Verification should be optional, situational, and accompanied by alternative pathways and appeals to avoid excluding vulnerable users.

4.4

Identifying and Flagging Harmful Content

For many participants, reporting was one of the few immediate actions available when harmful content appeared online. They described submitting reports when they encountered harmful material themselves or noticed that it affected others, particularly in cases involving intimate or explicit content (Scenarios G, C, D and I), impersonation (Scenarios B, D and F), or hateful and abusive posts and comments (Scenarios E, K and L) (see full scenario descriptions in Section Scenarios). However, existing reporting tools often felt slow, opaque, or disconnected from the outcomes. Users rarely received feedback and were uncertain whether anything changed as a result, leaving them unsure whether their efforts made a difference. As a result, participants expressed interest in approaches that go beyond formal reporting, suggesting lighter or more visible ways to flag content, signal concerns, and collectively surface patterns of harm.

Beyond simply removing content, participants described broader expectations for what flagging should achieve. Many saw it as a way to collectively surface patterns of harm and support safer communities. They wanted their actions to contribute meaningfully to prevention, for example by helping platforms better recognise abusive behaviour through AI systems, or by making it clearer how others identified and responded to the same content. In this sense, flagging was framed not only as an individual action, but also as a social signal that could inform moderation, improve detection, and build trust in the platform's response. Improvements and new features users asked for includes:

- Bystander flagging for explicit / impersonating content (widespread request)
- AI-driven detection and tracking of re-uploaded intimate media (common request)
- AI-based detection of harmful and culturally nuanced text (widespread request)

*Widespread: almost everyone said this, common: a lot of people said this, occasional: some mentioned, isolated: one-off idea.

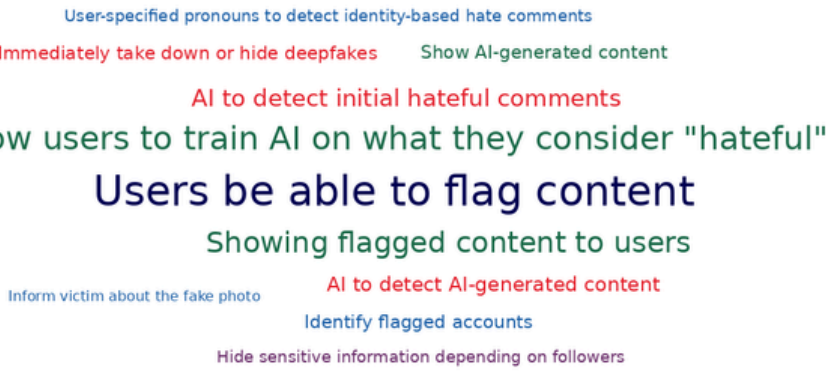


Figure 18: Key themes and ideas proposed by participants for identifying, flagging, and detecting harmful content on social media platforms.

4.4.1 Bystander flagging for explicit or impersonating content

Participants highlighted that intimate images, impersonating profiles, or manipulated content may circulate without the affected person immediately knowing. In such cases, friends, followers, or casual viewers may encounter the material first. However, existing systems place the burden of reporting primarily on the targeted individual. Participants therefore called for clearer mechanisms that allow bystanders to flag harmful content discreetly and responsibly, while ensuring the person affected is informed in a sensitive manner. They also expressed interest in making prior flagging visible, so others do not unknowingly amplify harmful material.

Participants described both supportive reporting pathways and sensitive warning cues:

"Report anonymously button on Instagram, if her friend reports it, [name] can receive a notification saying someone you know has reported this image, like it may be a harmful image of you. Before that to happen, I guess there's that question of how your friend would be able to see it. So I think there should be a feature where instead of the normal content you consume, Instagram prioritises people you may know above random content. And that's how people who know you can see it."

25-year-old Asian woman, active social media user who creates and posts content, consumes and reshares content, and participates in discussions

Another participant emphasised the importance of signalling when content has already been flagged:

"As a passerby, I want the platform to tell me if a video has been flagged already and why, so that I can feel more confident about not causing further spread or engagement with problematic content, like the current 'this video is done by professionals' tag shown at the bottom as a disclaimer."

23-year-old Asian woman, active social media user who creates and posts content, consumes content, and participates in discussions

EXPERT CONSIDERATIONS

Experts supported bystander flagging as a burden-reduction mechanism, but stressed that it needs strong guardrails: clear definitions, safe/trauma-aware notification pathways, and safeguards against misuse (for example, weighting flags based on account trust and minimising opportunities for false takedowns).

Effectiveness Experts emphasised that bystander flagging can reduce burden on victim-survivors, but only if there is clear guidance on what qualifies as 'harmful' and how flags are actioned.

Strengths and risks

Experts warned that the category 'harmful' can be subjective, creating risks of misuse, disagreement between bystanders, or takedowns that remove content the person may actually want online. Anonymous flagging can protect flaggers, but anonymity may also complicate accountability.

Practicality

Platforms may need ways to assess the 'trustworthiness' of the flagging account (for example, established accounts vs. disposable lurker accounts) so that flags can be weighted appropriately.

Policy and ethics

Experts raised concerns that notifying a victim that content exists 'about them' could be triggering, and needs careful and sensitive handling to avoid exacerbating harm.

Impact

Experts positioned bystander support as valuable particularly when the person targeted may be overwhelmed, avoiding content, or unable to safely engage with it directly.

4.4.2 AI-driven detection and tracking of re-uploaded intimate media

Participants described frustration at the need to repeatedly report the same intimate or explicit content when it was re-uploaded, slightly edited, or circulated through new accounts. They felt that once content had been identified as harmful, the platform should proactively prevent it from resurfacing. Rather than placing the burden on individuals to continuously search for and report copies, participants expected AI systems to detect patterns, recognise previously flagged media, and intervene early, ideally before the content becomes visible again.

"I want AI to spot my photos if they pop up anywhere, even edited, so that I don't have to keep finding this stuff on my own."

20-year-old Asian genderfluid person, active social media user who creates and posts content, consumes and reshares content, participates in discussions, and moderates social media accounts

One participant suggested a pattern to detect content posted with intent to harm an ex-partner:

"The idea is to make the ex unable to post in the first place. So you're trying to detect the words used in the captions, who's being tagged, and maybe compare the photo that's being posted with profile photos or known images. Then you combine all of that and detect whether it's a revenge post or not. That means you need a dataset of revenge posts, and once it's detected, it should stop the upload immediately."

34-year-old Asian woman, active social media user who consumes content

EXPERT CONSIDERATIONS

Experts saw strong potential in hashing/provenance-style detection to reduce repeated reporting and limit re-uploads, but flagged major implementation constraints: edited variants often evade matching, cross-platform coordination is limited, and governance/privacy questions (who runs the database, how victims are protected) are non-trivial.

Effectiveness

AI-based hashing or media-matching systems can reduce repeated victim reporting by identifying identical or near-identical re-uploads. However, edited variants such as cropped, filtered, or text-overlay versions may evade simple matching systems.

Strengths and risks

While automated detection can significantly reduce the re-circulation of harmful media, experts cautioned that over-reliance on automated systems may produce false positives or false negatives, particularly where contextual interpretation is required.

Practicality

Technical infrastructures for image hashing and database matching already exist in certain domains. However, scaling such systems requires platform investment, collaboration, and robust governance mechanisms.

Policy and ethics

Governance questions remain central, including who manages flagged-content databases, how victims' images are stored without redistributing harm, and how consent and privacy are preserved.

Impact

Experts emphasised that such systems align with the principle that victims should not bear the burden of continuously monitoring the internet for re-uploads. However, implementation must avoid creating additional pressure for victims to submit or retain sensitive media (for example, their biometric profile) in order to activate protection.

4.4.3 AI-based detection of harmful and culturally nuanced text

While many platforms already provide keyword filters or 'offensive comment' settings, participants expressed concern that these tools often fail to capture the more subtle, contextual, or culturally specific forms of harm they encounter. Comments are not always overtly aggressive; they can be sarcastic, backhanded, manipulative, or coded in ways that evade simple word-based filtering. Participants described frustration with block lists that rely on static keywords, noting that slang, alternative spellings, emojis, and cultural references frequently bypass automated systems. As a result, harmful comments can remain visible, while neutral or reclaimed language may be incorrectly flagged.

Some participants suggested that AI systems could move beyond fixed word lists toward sentiment analysis or contextual interpretation, potentially allowing users to actively teach the system what they personally consider harmful. However, they also questioned the feasibility and accuracy of such approaches.

"If for a comment there's AI like sentiment analysis and things like that... I was wondering if there is an option where the user dislikes a comment and it comes up, do you want to report the comment or delete the post, and that will help AI train on that and rank the comment. I'm not sure how feasible it is, but it might be helpful."

26-year-old Asian woman, social media user who consumes and reshares content

"Every country has different slang. It's hard to really know what words to filter, because that's also sometimes what happens in the comments."

26-year-old Asian woman, active social media user who creates and posts content, consumes and reshares content

EXPERT CONSIDERATIONS

Experts agreed that text-based harm detection needs to move beyond keywords toward context, but cautioned that cultural meaning shifts fast and 'intent' is hard to infer, so platforms likely need hybrid approaches (combination of context-sensitive cues, human review and strong appeal pathways) to avoid bias and over-blocking.

Effectiveness Detection failures often stem from reviewers or automated systems lacking sufficient cultural context, meaning reports may be dismissed as not hateful when the harm is culturally specific. User feedback mechanisms could help surface this context, though careful governance would be needed to prevent misuse.

Strengths and risks Meaning and norms shift quickly, and the same term can be harmful in one context and acceptable in another, raising risks of over-censorship or uneven enforcement.

Practicality Experts questioned the feasibility of operationalising 'cultural slang' at scale because it would require moderators (or models) to reference rapidly changing, context-dependent language knowledge.

Impact Experts supported automatic detection that can help users take action without having to read harmful content directly.

4.5 Tracking Abusers and Abusive Behaviours

This theme captures participants' interest in shifting from one-off responses to pattern-based monitoring of harmful behaviour. Rather than treating each incident in isolation, users wanted platforms to detect repeat offenders, connect related reports, and provide signals that help individuals recognise and respond to ongoing abuse more proactively.

Figure 19: Key themes and ideas proposed by participants for detecting patterns of abusive behaviour and monitoring suspicious account activity.

The following features were proposed:

- Risk signals from account history and reporting patterns (occasional request)
- Suspicious-activity insights and profile-visit pattern alerts (occasional request)

*Widespread: almost everyone said this, common: a lot of people said this, occasional: some mentioned, isolated: one-off idea.

4.5.1 Risk signals from account history and reporting patterns

Some participants wanted platforms to surface lightweight risk signals based on an account's prior behaviour (for example, how often the account has been reported, or whether it has been linked to previously flagged accounts). They described this as a way to help people make safer decisions before engaging (such as following someone new or replying), and as a way for platforms to take escalation more seriously when patterns are visible across multiple reports.

Participants felt that harms are difficult to prevent because warning signs are not visible until after something happens, and because reports are treated as one-off events rather than connected patterns across time and targets. They suggested that seeing whether an account had been repeatedly reported (and for what) could help them avoid further connection, and help platforms prioritise action:

"we can check the history whether this account was reported by many people or not so it can be the reference for the platforms to do something to this account..."

32-year-old Asian woman, active social media user who creates and posts content, consumes and reshares content

EXPERT CONSIDERATIONS

Experts generally aligned with users that platforms should connect reports into behavioural patterns, so action does not depend on repeated victim effort. However, they emphasised that any risk signal must be context-aware and governable, otherwise it can be misused or produce unfair outcomes.

Effectiveness Experts supported the idea of connecting reports over time to identify repeat offenders, but warned that counts alone can be misleading without context (for example, coordinated false reporting).

Strengths and risks Risk signals may deter abuse and support safer decisions, but can be weaponised (brigading, reputation damage) or lead to unfair targeting if transparency is poorly designed.

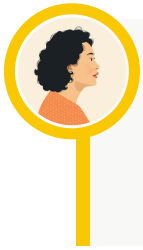
Practicality Requires better backend linking of reports across incidents (and potentially across linked accounts/devices), plus clear thresholds and review pathways so the system remains manageable.

Policy and ethics Experts raised governance questions about due process: what gets shown to whom, what evidence is required, how disputes/appeals work, and how to avoid defamation-like harms.

4.5.2 Suspicious-activity insights and profile-visit pattern alerts

Participants also wanted access to simple analytics to help them detect stalking-like patterns (for example, repeated profile visits, repeated views from new accounts, or unusual activity linked to the same person). This was often framed as a way to regain a sense of control, especially when abusers could still observe them via fake accounts or mutual connections.

Users described feeling unable to prevent escalation because they could not see who was repeatedly monitoring them, particularly after blocking. They wanted a way to identify patterns (rather than manually guessing) and receive proactive alerts:



"If suspicious activity is detected... repeated views from new accounts... trigger a private alert... one tap to block, restrict or report."

36-year-old Asian woman, active social media user who consumes content and participates in discussions

EXPERT CONSIDERATIONS

Experts saw value in pattern-based alerts that help users act early (block, restrict interactions or report) without intensive self-monitoring. However, they stressed that analytics must be privacy-preserving, thresholded, and safety-centred, otherwise it risks becoming a surveillance feature or increasing anxiety.

Effectiveness	Experts noted these signals can help users recognise patterns earlier, but 'view counts' do not always equal harm (and can create false alarms).
Strengths and risks	Strong potential for reassurance and control, but high privacy risk: it may reveal information about viewers, enable retaliation, or be exploited by abusers to confirm they are being watched.
Practicality	Implementation is technically feasible, but needs careful defaults, thresholds, and safe actions (for example, 'tap to restrict') so it does not become noisy or overwhelming.

4.6 Training Users on Safety Features and Safety Awareness

Across all workshops, participants expressed uncertainty about how to navigate existing safety features during moments of distress. Many described searching Google or LLM-based (Large Language Model) tools like ChatGPT, asking friends, or navigating complex help centres when facing harassment, impersonation, or privacy concerns. Even when safety tools existed, they were often difficult to locate, poorly explained, or buried within changing interface layouts.

Participants also highlighted that safety awareness is not just about crisis response. Many felt platforms do little to proactively educate users about privacy controls, tagging settings, data visibility, or emerging risks. They wanted clearer onboarding, contextual prompts, and ongoing guidance that supports prevention, not just reactive reporting.

Features and improvements users asked for:

- AI-based guidance or safety assistants (common request)
- Improved onboarding and contextual prompts (widespread request)

*Widespread: almost everyone said this, common: a lot of people said this, occasional: some mentioned, isolated: one-off idea.

Figure 20: Key themes and ideas proposed by participants for improving the discoverability and accessibility of safety features on social media platforms.

4.6.1 AI-powered guidance and real-time safety assistants

Many participants described feeling overwhelmed when trying to respond to harmful situations. They often did not know which feature to use, where it was located, or what settings would meaningfully reduce harm. Many noted that they would first Google the issue or ask friends, rather than rely on the platform itself. Searching externally for answers increased anxiety and delayed action.

Participants proposed a built-in AI assistant or chatbot that could:

- Allow users to describe their situation in plain language
- Suggest relevant safety features in context
- Provide step-by-step guidance
- Offer real-time support
- Escalate to human assistance if needed

One participant framed it as:

> *"I want to get real-time, fast support from the platform (for example, through a 24/7 chatbot) on what features or steps I can take, in the event of an abusive situation."*
>
> 34-year-old Asian woman, social media user who consumes and reshares others' content

EXPERT CONSIDERATIONS

Experts viewed AI-based safety assistants as potentially useful for navigation and early-stage guidance, provided they are carefully constrained and integrated with human oversight. Without strong guardrails, bias mitigation, and clear duty-of-care protocols, such systems risk misuse or unintended harm.

Effectiveness	AI-based guidance could improve discoverability of safety tools and reduce friction in navigating complex interfaces.
Strengths and risks	Experts raised concerns about misuse, including perpetrators reframing harmful intentions as 'safety' queries. There were also concerns about AI reproducing bias, misunderstanding context, or responding in ways that could retraumatise vulnerable users.
Practicality	AI should redirect users to verified resources and controls rather than independently interpret or generate sensitive advice. Guardrails and clear boundaries would be essential.
Policy and ethics	Duty of care and data governance were key concerns. Experts questioned who stores user disclosures, how queries are logged, and what happens if users reveal high-risk situations (for example, coercive control, suicidal ideation). Regional and legal variations were also highlighted.

4.6.2 Improved onboarding and contextual prompts

Across workshops, participants consistently described not knowing what safety features exist, where to find them, or how they work. Many participants reflected that safety controls are frequently buried within settings, vaguely labelled, or introduced without explanation. Frequent interface updates meant users missed newly implemented protections. In moments of distress, trying to search through menus or interpret technical terms added to anxiety rather than reducing it.

As one participant explained:

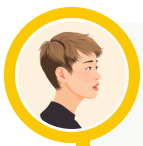
> *"I think there needs to be greater feature awareness and promotion when people open accounts... Don't just say account setting X, what does this mean, what does it do?"*
>
> 60-year-old European transgender woman, active social media user who creates and posts content, consumes and reshares content, participates in discussions, and moderates social media accounts

Participants proposed embedding safety learning directly into the platform experience rather than isolating it in help centres. Two main approaches were repeatedly suggested:

- Proactive onboarding
  - Integrating safety setup during account creation
  - Clear explanations of what each setting does and why it matters
  - Guided configuration of tagging, direct message permissions, and blocking tools
  - Periodic 'safety check-ups' to check users' current safety settings and reevaluate rather than one-off setup

- Contextual prompts and smart nudges
  - Pop-ups triggered when unusual or harmful activity is detected
  - AI prompts suggesting protective actions (for example, "Would you like to restrict this account? Learn more about restricting here.")
  - Alerts when repeated tagging or commenting patterns are identified
  - Quick, one-click navigation to relevant privacy and safety settings

For example:



> *"Once a lot of comments start popping up in an unusual way, the message would come up... would you like to take any action?"*
>
> 20-year-old Asian non-binary person, mostly consumes content

EXPERT CONSIDERATIONS

Experts broadly supported integrating safety education into onboarding and contextual prompts, but stressed careful pacing, clarity, and inclusivity. Well-designed nudges may strengthen user awareness and agency, whereas poorly timed or overly frequent prompts risk disengagement.

Effectiveness

Experts agreed that proactive safety education can reduce accidental exposure and improve user agency, particularly around privacy and visibility controls.

Practicality

Experts cautioned against cognitive overload. Dense onboarding flows may lead to users skipping content without understanding it. They suggested progressive disclosure, staged reminders, and context-sensitive prompts rather than front-loading all safety information at sign-up.

Impact

Experts emphasised the importance of plain language, multilingual delivery, and culturally sensitive examples of harmful scenarios and safety responses. Without this, safety education risks benefiting only digitally confident users while excluding vulnerable groups.

Section 5

Discussion

This project examined how women and gender-diverse social media users experience and navigate platform safety features in the context of technology-facilitated abuse. Through co-design workshops with users and expert consultations, participants reflected on everyday platform interactions, reporting processes, content sharing practices, identity verification, and mechanisms for detecting or preventing abuse. Rather than focusing solely on the availability of safety features, participants' discussions revealed deeper tensions between how platforms conceptualise harm and how harm is experienced in practice.

Key Insight

Online harm cannot be defined by content alone

A highly recurring theme observed was that the definition of harm on social media platforms is often unclear, incomplete, or poorly aligned with users' lived experiences. Participants frequently described situations where harmful interactions were difficult to categorise within existing reporting frameworks or moderation guidelines. What constitutes harm, participants suggested, cannot always be determined from the content of a message or post alone. Instead, harm is often shaped by relationship history [17, 21], cultural meaning, patterns of behaviour, and broader social context [30].

For example, messages that may appear neutral in isolation, such as "Let's meet up", can become threatening depending on prior interactions between individuals. Similarly, comments that target cultural identity, religion, or gender may carry meanings that automated moderation systems or standard reporting categories fail to capture. Participants also described how engagement algorithms, cross-platform harassment, and repeated behavioural patterns can intensify the experience of harm even when individual pieces of content do not appear to violate platform rules.

These findings highlight a fundamental gap between platform moderation models, which often operate at the level of individual pieces of content, and users' experiences of abuse, which are relational and cumulative. When harm is evaluated primarily through content-level signals, the broader context that gives interactions their meaning may remain invisible to detection systems or reporting processes [16]. As a result, users often feel that moderation systems fail to recognise the seriousness of their experiences or respond appropriately.

The difficulty of defining harm also influences how users engage with safety mechanisms. Participants described hesitation in reporting certain incidents because they were unsure whether the behaviour would be recognised as abuse, or because previous reports had been dismissed as not violating platform policies [24]. In this way, uncertainty about how harm is defined can discourage reporting and contribute to the normalisation of harmful behaviour within online environments [12].

Key Insight

Opaque reporting systems undermine user trust

Participants emphasised that reporting systems often feel opaque and difficult to understand. While most platforms provide reporting options, users described uncertainty about what happens after a report is submitted. Many said they receive little or no information about whether their report was reviewed, who reviewed it, or what criteria were used to determine the outcome. This lack of visibility made it difficult for participants to feel confident that their concerns were being taken seriously [8].

Participants did not simply want visual indicators such as progress bars or automated confirmations. Instead, they wanted clearer explanations about how moderation decisions are made. This included understanding whether reports are reviewed by automated systems or human moderators, how cases are prioritised, and whether multiple reports about the same person or behaviour are considered together. Without this information, participants often interpreted delayed or absent responses as inaction or dismissal.

Transparency therefore emerged as an important factor in building trust in reporting systems. When users cannot see how decisions are made or why certain actions are taken, moderation processes can appear arbitrary or unresponsive. These concerns echo broader Safety by Design principles, which emphasise the importance of transparency and clear accountability in how platforms manage online harms [10].

A lack of transparency may also discourage people from reporting harm in the first place. Previous research has similarly shown that users may choose not to report abuse if they believe nothing will change or if past reports have not resulted in visible action [12]. In this way, transparency is not only a design issue but also a governance signal that communicates whether platforms are responsive to users' concerns and committed to addressing harm.

Key Insight

Preventing abuse through friction while avoiding exclusion

Both users and experts discussed preventative approaches that introduce friction into platform interactions in order to reduce abuse. Participants often described the ease with which harmful actions can occur online, such as creating multiple accounts, sending unwanted messages, or uploading sensitive images involving other people. As a result, many users supported measures that would slow down or interrupt these actions before harm occurs.

Some suggestions focused on platform governance-level measures, such as stronger identity verification processes, limits on creating multiple accounts, or mechanisms that make it harder for people who have been blocked or banned to reappear under new identities. Participants felt that these measures could improve accountability and reduce repeat harassment.

Other suggestions involved design-level friction embedded directly into everyday platform interactions. Examples included warning prompts when uploading images that contain other people's faces, consent checks before posting intimate or sensitive content, or reminders encouraging users to reconsider messages that may be harmful. Participants viewed these

small interruptions as opportunities for users to pause and reflect before posting or sharing potentially abusive content, consistent with prior co-design research demonstrating that friction-based interventions can be effective in reducing harmful sharing behaviours [1].

However, both users and experts also recognised that increasing friction can introduce new risks. Measures such as strict identity verification or limits on multiple accounts may disproportionately affect vulnerable groups who rely on anonymity or pseudonyms for safety, including survivors of abuse, LGBTIQ+ users [2], activists [20], and people participating in sensitive online communities [5]. In these cases, stronger identification requirements may unintentionally expose individuals or discourage them from seeking support online, and in some jurisdictions mandatory real-name policies have been found to be unlawful [19].

This tension highlights the challenge platforms face in designing preventative measures. While introducing friction may help reduce abusive behaviour, poorly designed restrictions may also create barriers for legitimate users. As a result, participants emphasised that preventative features should be carefully designed to balance accountability with accessibility and safety for vulnerable communities.

Key Insight

Predicting harm and surfacing safety features at the right moment

Participants frequently discussed the potential role of artificial intelligence and behavioural analysis in identifying patterns of abuse before harm escalates. Many users described online abuse as ongoing and patterned rather than isolated incidents. For example, harassment may involve repeated messages, the creation of multiple accounts after being blocked, persistent commenting across posts, or the circulation of explicit or fabricated images over time. Because of this, participants expected platforms to look beyond individual pieces of content and recognise behavioural patterns that signal possible abuse.

Several participants noted that social media platforms already collect large amounts of behavioural data about user activity [32]. As a result, they questioned why these insights are not used more proactively to identify harmful patterns such as repeated harassment, stalking behaviours, coordinated attacks, or the repeated sharing of intimate images without consent. From this perspective, predictive detection was seen as a potential way to intervene earlier, before harm escalates. This aligns with prior participatory work showing that contextual prompts (‘nudges’) and clearer system explanations can support users in understanding platform behaviour and managing privacy or safety risks [26].

Experts generally agreed that technological advances make this form of pattern recognition increasingly feasible. However, they also emphasised the need for caution. Systems designed to predict abuse may misinterpret behaviour, incorrectly flag legitimate interactions, or introduce new forms of surveillance if not carefully governed. Experts also raised broader questions about how harm is defined within predictive systems and whose perspectives shape those definitions. As noted in recent work by the eSafety Commissioner, the lived experiences of people involved in training these systems are important to ensure that culturally specific and contextual forms of harm are recognised and addressed appropriately [25].

Participants also connected predictive technologies with another concern: the visibility and learnability of safety features. Many described safety tools such as reporting, blocking, filtering, or privacy controls as difficult to find or buried within complex settings pages. Rather than expecting users to search for these tools after harm occurs, participants suggested that platforms could surface relevant safety features when patterns of risk are detected. For example, contextual prompts could remind users about blocking options, privacy controls, or reporting pathways at moments when harmful interactions are likely, or surface lesser-known features such as interaction limits when more familiar actions like blocking are taken.

This approach aligns with Safety by Design principles that encourage platforms to support user autonomy by making safety tools visible and accessible at the point where they are most relevant [9,10]. When implemented carefully, predictive systems may therefore serve not only to detect potential harm but also to guide users toward protective actions before situations escalate.

Key Insight

Platform safety as an ongoing governance responsibility

While many of the discussions in this report focus on specific features and design improvements, participants and experts repeatedly emphasised that platform safety cannot be addressed through technical tools alone. Decisions about what constitutes harm, how reports are prioritised, and when interventions occur are ultimately questions of governance. Safety by Design frameworks similarly emphasise that addressing technology-facilitated abuse requires organisational accountability and governance mechanisms beyond individual platform features [10].

Participants highlighted the importance of involving diverse communities in shaping these definitions. Experiences of online harm often differ across cultural groups, genders, and communities, and static rule sets may fail to capture these differences [6, 28]. Users and experts suggested that platforms should treat safety as an ongoing conversation rather than a fixed set of rules, creating mechanisms for users, researchers, and community organisations to contribute to how harms are identified, interpreted, and addressed over time.

Experts also noted that platforms cannot address technology-facilitated abuse in isolation. Many forms of harm intersect with legal systems, law enforcement, and support services that operate outside the platform itself. As a result, stronger coordination between social media companies, regulators, and specialist support organisations may be necessary. This could include clearer pathways for escalating serious cases, partnerships with victim-support services, and processes that help connect affected users with local resources that are better equipped to respond to complex safety situations.

Strengthening social media safety requires both design improvements and broader governance approaches. Platforms may play a critical role in shaping safer online environments, but effective responses to technology-facilitated abuse will likely depend on ongoing collaboration between platforms, policymakers, researchers, and community organisations. Framing safety as a shared and evolving responsibility may help ensure that responses remain responsive to changing technologies, emerging harms, and the lived experiences of those most affected.

Research Team

Senuri Wijenayake

Dr Senuri Wijenayake is a Senior Lecturer and an ARC DECRA Fellow in the School of Computing Technologies at RMIT University. She is a Human–Computer Interaction (HCI) researcher whose work examines online harms on social media platforms, focusing on the types of harms people experience, how users respond to them, and how platform safety features are used in these processes. Her research particularly centres on the online safety needs and experiences of under-represented populations at greater risk of harm, including women, gender diverse users, young people, and culturally and linguistically diverse communities. Her work integrates perspectives from computing and the social sciences to inform the design and governance of safer digital systems, including Safety by Design approaches and platform policies that support both preventative and responsive mechanisms for online safety.

Madhuka De Silva

Madhuka De Silva is a researcher in the School of Computing Technologies at RMIT University. She holds a PhD in Human-Centred Computing from Monash University. Her work sits at the intersection of co-design, digital inclusion, and online safety, with a focus on how participatory and community-engaged methods can surface the needs and experiences of under-represented groups in digital contexts. She has extensive experience facilitating co-design workshops with community participants and sector experts, and applying mixed-methods approaches to translate lived experience into evidence-based design and policy directions.

Dana McKay

Dana McKay is an Associate Professor and Associate Dean of the Interaction, Technology and Information Discipline in the School of Computing Technologies at RMIT. Her work sits at the nexus of Information Science and Human-Computer Interaction. Her research focus is on how advances in information technologies affect minoritised groups, and how the properties of these advances might be harnessed to increase social justice and equity. To that end, she studies the impact of technology platforms on information flows, how people find, manage, use, and abuse information, how people can control information about themselves, and how information can be provided in emancipatory ways to minoritised groups.

Anastasia Powell

Anastasia Powell is Professor of Family \& Sexual Violence, in Criminology & Justice Studies at RMIT. Anastasia has over 20 years of experience as a criminologist in the fields of prevention, response and recovery, addressing family and sexual violence. Her recent research has examined: sexual violence prevention practice; sector responses to family and domestic violence; the intersections of gender-based violence and technology-facilitated abuse; and cyber crimes. She is the author or co-author of over 100 scholarly works including books such

as: The Palgrave Handbook on Gendered Violence and Technology (2021), Digital Criminology (2018), and Sexual Violence in a Digital Age (2017). Anastasia brings together evidence-based expertise, as well as her own lived experience of violence, to her research, teaching and advocacy work.

Asangi Jayatilaka

Dr Asangi Jayatilaka is a cyber security and software engineering lecturer in the School of Computing Technologies at RMIT University. She has extensive expertise in human-centric cyber security and has co-authored several publications in top-tier venues. Dr Jayatilaka has a strong track record in applying participatory design methodologies, including co-design, to investigate and improve digital technology experiences among marginalised communities such as women, gender diverse people, and people with cognitive impairments. Through this work, she has developed practical, evidence-based insights into how security solutions can be designed to be more inclusive, usable and effective in real-world contexts. She has led multiple Cyber Security Cooperative Research Centre (Cyber Security CRC)-funded projects on human-centric cyber security, spanning areas such as secure behaviour, security awareness and education, and the development of user-centred security tools and frameworks.

Danula Hettiachchi

Danula Hettiachchi is a Lecturer in the School of Computing Technologies at RMIT University and an Associate Investigator at the ARC Centre of Excellence for Automated Decision-Making and Society (ADM+S). His research focuses on human-computer interaction, social computing, crowdsourcing, and responsible AI, with a particular interest in designing fair and transparent systems for human-AI collaboration. His research has attracted international funding, including the 2025 Google Research Scholar Award. He brings extensive practical experience from prior industry roles in machine learning, data engineering, and software development, complementing his strong background in quantitative user studies.

Tuck Wah Leong

Tuck Wah Leong is a Professor at the School of Computing Technologies, RMIT who specialises in Human–Computer Interaction and Interaction Design. His research explores how digital technologies can be designed more thoughtfully, inclusively, and meaningfully for everyday life. His work centres on underserved and under-represented communities, including Aboriginal and Torres Strait Islander communities, older adults, pregnant women and new mothers seeking culturally safe care, adults living with early onset dementia, queer youth, and families negotiating technology use in the home. Drawing on participatory design, experience-centred design, and other human-centred approaches, he investigates how values, lived experience, and social context can shape better technologies and more respectful design practices.

Joanne E. Gray

Dr Joanne E. Gray is Chair of Media and Communications at the University of Sydney and an internationally recognised scholar of digital technology policy and governance. She is Editor-in-Chief of the Q1 journal Policy & Internet, Lead Investigator of the Governing Immersive Tech

Project (ARC DP25), and Chief Investigator on the For You Project (ARC DP24). Across her research and teaching, Gray's work is designed to improve understanding of how governments, platforms, and societies can identify and proactively manage the risks and opportunities of digital technologies, from social media to AI and VR/AR.

Luke Hespanhol

Associate Professor Luke Hespanhol is the Head of Discipline (Design) at the School of Architecture, Design and Planning, The University of Sydney. His research focus on people, culture, and technology, including social design, civic participation, storytelling, cities, digital placemaking, technology-mediated social interactions, and cross-cultural education. An acknowledged global leader in his field, Luke has pioneered mobility programs in Design, with particular focus on the Indo-Pacific and multiple programs in China and Indonesia. In Australia, Luke has contributed to translation of research into public discourse by leading regular seminars and public events on the societal impacts of technology and how to design more sustainable futures. In addition to his contributions to Design, Luke holds a certificate in Dispute Resolution by the Program on Negotiation at Harvard Law School, USA, and a fellowship from the Social Impact Hub, Australia.

Justine Humphry

Justine Humphry is an Associate Professor of Digital Cultures in Media and Communications at the School of Art, Communication and English, The University of Sydney. Her research is internationally recognised for its critical focus on how mobile and social media are embedded in everyday life and shape experiences of social inclusion and exclusion. Justine's research on emerging online safety issues funded by the eSafety Commission used co-design and co-creative methods to involve young people and parents, in producing evidence-based social media education addressing new kinds of algorithmic and data-related online harms. Her prior on work the use of mobile phones among people experiencing homelessness also examines the essential role of mobile connectivity in accessing services, maintaining social networks and navigating urban space, documented in her book: Homelessness and Mobile Communication – Precariously Connected. Justine works closely with communities and social service providers to build capacity and engage with public and policy debates on digital inclusion and the social implications of emerging technologies, shaping conversations about more equitable and inclusive digital futures.

Anjalee de Silva

Anjalee de Silva is a Senior Lecturer at Melbourne Law School, The University of Melbourne, an Associate Investigator at the ARC Centre of Excellence for Automated Decision-Making and Society, and a Women's Leadership Institute Australia Fellow. She is an expert in administrative, anti-discrimination, and free speech and media law, with a focus on harmful speech and its regulation, especially as it relates to women and in online contexts. She has an upcoming book with Cambridge University Press, Hate Speech Against Women and the Role of Law.

References

[1] Zainab Agha, Karla Badillo-Urquiola, and Pamela J. Wisniewski. 2023. "Strike at the Root": Co-designing Real-Time Social Media Interventions for Adolescent Online Risk Prevention. Proc. ACM Hum.-Comput. Interact. 7, CSCW1, Article 149 (April 2023), 32 pages. doi: 10.1145/3579625

[2] Tommaso Armstrong and Tuck Wah Leong. 2020. SNS and the Lived Experiences of Queer Youth. In Proceedings of the 31st Australian Conference on Human-Computer-Interaction (Fremantle, WA, Australia) (OzCHI'19). Association for Computing Machinery, New York, NY, USA, 376380. doi:10.1145/3369457.3369497

[3] Australian Bureau of Statistics. 2022. 2021 Census: Nearly half of Australians have a parent born overseas. https://www.abs.gov.au/media-centre/media-releases/2021-census-nearly-half-australians-have-parent-born-overseas. Media release, Australian Bureau of Statistics. Accessed 21 February 2026

[4] Australian Bureau of Statistics. 2022. Cultural diversity of Australia. https://www.abs.gov.au/articles/cultural-diversity-australia. Article released 20 September 2022 on the Australian Bureau of Statistics website. Accessed 21 February 2026.

[5] Amna Batool, Mustafa Naseem, and Kentaro Toyama. 2024. Expanding Concepts of Non-Consensual Image-Disclosure Abuse: A Study of NCIDA in Pakistan. In Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems (Honolulu, HI, USA) (CHI '24). Association for Computing Machinery, New York, NY, USA, Article 398, 17 pages. doi:10.1145/3613904.3642871

[6] Cynthia Brown, Michael Flood, and Kelsey Hegarty. 2022. Digital dating abuse perpetration and impact: The importance of gender. Journal of Youth Studies 25, 2 (2022), 193–208.

[7] eSafety Commissioner. 2017. Image-based Abuse: National Survey-Summary Report. https://www.esafety.gov.au/sites/default/files/2019-07/Image-based-abuse-national-survey-summary-report-2017.pdf

[8] eSafety Commissioner. 2023. Australians negative online experiences 2022. https://www.esafety.gov.au/research/australians-negative-online-experiences-2022

[9] eSafety Commissioner. 2024. Safety by Design. https://www.esafety.gov.au/industry/safety-by-design

[10] eSafety Commissioner. 2024. Technology, gendered violence and Safety by Design. Industry Guide. eSafety Commissioner, Australian Government. https://www.esafety.gov.au/industry/safety-by-design/industry-guides Safety by Design industry guide. Accessed 5 March 2026.

[11] eSafety Commissioner. 2024. Women In The Spotlight: How Online Abuse Impacts Women in Their Working Lives. https://www.esafety.gov.au/research/how-online-abuse-impacts-women-working-lives/report

[12] eSafety Commissioner. 2025. Fighting the tide: Encounters with online hate among targeted groups. https://www.esafety.gov.au/research/encounters-with-online-hate

[13] eSafety Commissioner. 2025. Online risks for women. https://www.esafety.gov.au/women/online-risks-for-women

[14] eSafety Commissioner. n.d.. Encounters with online hate. https://www.esafety.gov.au/research/encounters-with-online-hate. eSafety Commissioner research page on encounters with online hate. Accessed 21 February 2026

[15] Asher Flynn, Sophie Hindes, and Anastasia Powell. 2022. Technology-facilitated abuse: Interviews with victims and survivors and perpetrators. Australia's National Research Organisation for Women's Safety (ANROWS).

[16] Asher Flynn, Lisa J Wheildon, Brady Robards, Zarina Vakhitova, and Bridget A Harris. 2023. Australian users experiences with control features on social media services and online dating apps. (2023).
[17] Nicola Henry, Clare McGlynn, Asher Flynn, Kelly Johnson, Anastasia Powell, and Adrian J Scott. 2020. Image-based sexual abuse: A study on the causes and consequences of non-consensual nude or sexual imagery. Routledge, United Kingdom.
[18] Nicola Henry and Anastasia Powell. 2018. Technology-facilitated sexual violence: A literature review of empirical research. Trauma, violence, & abuse 19, 2 (2018), 195–208.
[19] Matthew Humphries. 2018. German Court Rules Facebook's Real-Name Policy Is Illegal. https://au.pcmag.com/social-media/51753/german-court-rules-facebooks-real-name-policy-is-illegal Accessed: 24 March 2026.
[20] Anna Ricarda Luther, Hendrik Heuer, Stephanie Geise, Sebastian Haunss, and Andreas Breiter. 2025. Social Media for Activists: Reimagining Safety, Content Presentation, and Workflows. In Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25). Association for Computing Machinery, New York, NY, USA, Article 956, 18 pages. doi:10.1145/3706598.3713351
[21] Clare McGlynn, Erika Rackley, Kelly Johnson, Nicola Henry, Asher Flynn, Anastasia Powell, Nicola Gavey, and Adrian J. Scott. 2019. Shattering lives and myths: A report on image-based sexual abuse. Project Report. Durham University, University of Kent, RMIT University, Monash University, University of Auckland, Goldsmiths, University of London.
[22] Meta Platforms, Inc. n.d.. Confirm Your Identity. https://www.facebook.com/help/contact/515009838910929. Facebook Help Centre contact form. Accessed 30 January 2026.
[23] Meta Platforms, Inc. n.d.. Report a Facebook profile or Page pretending to be you or someone else. https://www.facebook.com/help/174210519303259. Facebook Help Centre. Accessed 30 January 2026.
[24] Office of the eSafety Commissioner. 2019. Women from Culturally and Linguistically Diverse Backgrounds:Summary Report. Technical Report. Australian Government. https://www.esafety.gov.au/research/women-from-diverse-backgrounds Accessed 2026.
[25] Office of the eSafety Commissioner. 2023. Tech Trends Position Statement: Generative AI. Technical Report. Australian Government. https://www.esafety.gov.au/ Position statement as of 15 August 2023.
[26] Jinkyung Katie Park, Renkai Ma, Naima Samreen Ali, Naulsberry Jean Baptiste, Zainab Agha, and Pamela J. Wisniewski. 2025. Teens, Privacy, and Algorithms: Navigating and Co-Designing Solutions for Interpersonal Boundary Management on Social Media. In Proceedings of the 24th Interaction Design and Children (IDC'25). Association for Computing Machinery, New York, NY, USA, 589607. doi:10.1145/3713043.3728840
[27] Jess Ringrose, Kate Regehr, and Ben Milne. 2021. Understanding and Combatting Youth Experiences of Image-Based Sexual Harassment and Abuse. Report. UCL Institute of Education, University College London. Open access report deposited in UCL Discovery. Accessed 23 February 2026.
[28] 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. 1–14.
[29] Morgan Klaus Scheuerman, Stacy M Branham, and Foad Hamidi. 2018. Safe spaces and safe places: Unpacking technology-mediated experiences of safety and harm with transgender people. Proceedings of the ACM on Human-computer Interaction 2, CSCW (2018), 1–27.
[30] Nicolas Suzor, Molly Dragiewicz, Bridget Harris, Rosalie Gillett, Jean Burgess, and Tess Van Geelen. 2019. Human rights by design: The responsibilities of social media platforms to address

gender-based violence online. Policy & internet 11, 1 (2019), 84–103.
[31] TikTok Pte. Ltd. n.d.. Web Account FAQ. https://www.tiktok.com/support/faq_detail?query=impersonation&id=7611812701194639883&category=web_account TikTok Support Centre FAQ page on web account issues. Accessed 30 January 2026.
[32] Shoshana Zuboff. 2023. The age of surveillance capitalism. In Social theory re-wired. Routledge, 203–213.
[33] Madeleine Janickyj and Leonie Maria Tanczer. 2025. Tech Abuse Personas: Exploring Help-Seeking Behaviours and Support Needs of Victim/Survivors of Technology-Facilitated Abuse. In Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA'25). Association for Computing Machinery, New York, NY, USA, Article 509, 11 pages. doi:10.1145/3706599.3719986
[34] Woodlock, D., Bentley, K., Schulze, D., Mahoney, N., Chung, D., and Pracilio, A., (2020). Second National Survey of Technology Abuse and Domestic Violence in Australia. WESNET.
[35] eSafety Commissioner. 2019. Adults' Negative Online Experiences. eSafety Commissioner, Canberra, Australia. Retrieved from https://www.esafety.gov.au/research/adults-negative-online-experiences

Appendix A: Scenarios

Scenario A
Persona: Sophie, 19 years old, woman

Sophie (She/her), a 19-year-old woman begins receiving late-night messages on [Facebook/ Instagram/ TikTok] from someone she does not know. At first, the messages are short and unusual, but over time they become persistent and uncomfortable, pressuring her to respond. Sophie ignores them, hoping they will stop, but the messages continue each night. The repeated contact makes her feel uneasy and concerned about her safety, leaving her unsure how to manage the ongoing intrusion.

Scenario B
Persona: Jordan, 37 years old, professional (sexually identified as LGBTIQ+)

Jordan (they/them), a 37-year-old professional, has a disagreement with a colleague at work. Shortly after, Jordan discovers a fake Instagram account using their name and profile photo. The account is sharing misleading content with captions and images that misrepresent Jordan, some of which are seen by mutual work contacts and spark unhelpful gossip. Although Jordan reports the account and it is eventually removed, the experience leaves them worried that more fake accounts could be created in the future.

Scenario C
Persona: Meera, 22 years old, international student

Meera (She/her), a 22-year-old international student who recently arrived in Australia, uses [Facebook/Instagram/TikTok] to make new connections. She begins chatting with a young man she doesn't know in real life, who initially seems friendly and supportive, often messaging her about university life and adjusting to a new country. Over the first few weeks, their conversations become more personal, and after repeated requests, Meera shares a few photos. Later, she starts feeling uncomfortable with the direction of the chats and tells him she doesn't want to send any more. His tone quickly changes, and he begins sending upset messages, attaching screenshots of the images she had previously shared, and hinting that he might show them to her classmates or family overseas unless she continues sending more.

Scenario D
Persona: Kara, 36 years old woman

Kara (She/her), a 36-year-old woman, notices her ex-partner has started making negative posts about her on Facebook. He uploads old photos from their relationship with captions such as "Can't believe I wasted years on this" and "Funny how loyalty means nothing to some". He also shares memes about betrayal and toxic relationships, tagging her so the posts appear in her notifications. Wanting to avoid the situation, Kara blocks him on Facebook. A few days later, he begins posting similar content on Instagram, this time with additional photos and mocking captions, again tagging her. When Kara blocks him there, he switches to TikTok, uploading short videos stitched with trending sounds and including screenshots from her old posts. On TikTok, he tags both Kara and some of her close friends, which makes the content more visible in her social circle.

Scenario E
Persona: Liam, 20 years old, gender diverse person

Liam (he/they/them), a 20-year-old gender diverse person, regularly shares photos on social media to stay connected with friends. Over time, one of his male friends begins leaving sarcastic comments on many of his posts — remarks about Liam's appearance, clothing, or sexuality. Some comments refer to his gender, such as "Settle down, princess" with laughing emojis. On other occasions, the friend makes replies that question Liam's identity or highlight past experiences in a way that draws unwanted attention. When Liam raises the issue, the friend dismisses it as light-hearted 'banter'. The behaviour continues, and sometimes the friend tags others to join in, which makes Liam feel increasingly singled out and hesitant to share content online.

Scenario F
Persona: Sofia, 30 years old, woman

Sofia (She/ her), a 30-year-old woman, notices unusual activity on her social media account a few weeks after ending a difficult relationship. New posts appear that she didn't create, some of her contacts are unexpectedly blocked, and inappropriate comments are sent to friends from her account. She realises her ex-partner may have gained access to her login details, which she had previously shared with him while they were together. Before long, he changes the recovery email and phone number, leaving her unable to log back in. Friends begin sending her screenshots of stories posted from her account, suggesting things about her personal life that aren't true and causing confusion among her social circle. Sofia wants to regain control of her account, prevent further misuse, and ensure her information stays secure in the future.

Scenario G
Persona: Olivia, 21 years old, woman

Olivia (She/her), a 21-year-old woman, is in a turbulent relationship with her boyfriend. During a disagreement, she discovers that he has shared intimate videos of her on a social media platform without her permission. When she asks him about it, he says it is a 'private post' but does not take it down. The content is visible online for several weeks, and by the time it is removed, it has already been seen and shared by others. Over time, he uploads similar content from new accounts or slightly altered versions, making it difficult for Olivia to feel that the situation is resolved. The repeated circulation leaves her feeling exposed, frustrated, and unsure how to regain control over her privacy.

Scenario H
Persona: Maria, 45 years old, female

Maria (She/her), a 45-year-old woman, starts receiving unwanted late-night messages from her ex-partner on Facebook and Instagram, pressuring her to meet up. Over the following days, new profiles appear with different names and profile pictures, and the messages become more persistent, sometimes referencing details about her daily routine that leave Maria feeling uneasy and unsettled. The repeated contact makes her feel watched and anxious, and she wants a way to prevent the same person from continuing to reach out to her online.

Scenario I
Persona: Isabella, 23 years old, woman

Isabella (She/her), a 23-year-old woman, discovers that altered images of her, created using AI, are appearing on a social media platform. A stranger has taken publicly available photos from her Instagram account and used them to generate realistic, sexualised images. Some posts tag her account, meaning that friends, colleagues, and family members may see them. Similar images continue to appear over time, leaving Isabella feeling exposed, distressed, and anxious about her online presence. She wants a way to feel that the images will no longer circulate and that her online privacy is protected.

Scenario J
Persona: Ella, 24 years old, woman

Ella (She/her), a 24-year-old woman, notices that her ex-partner continues to monitor her online activity months after their breakup. He regularly views her public posts on [Facebook/Instagram/TikTok] and engages just enough to make his presence noticeable — liking posts, reacting to stories, and leaving brief comments such as ‘thumbs up emoji’ or ‘Nice’. This behaviour feels deliberate, as if to remind her he is still watching. The ongoing monitoring and subtle interactions make Ella feel self-conscious and hesitant to post or engage online freely.

Scenario K
Persona: Maya, 25 years old, woman

Maya, a 25-year-old woman, uses Facebook to share posts about her cultural traditions and religious celebrations. After posting photos from a community festival, she begins receiving mocking comments from some users with opposing views. Initially, the comments focus on her appearance and clothing, but over time they include critical remarks about her religion and culture. Some group members create memes using her photos with captions that misrepresent her, and they share them in a public group where Maya is tagged. As the posts spread, strangers also send private messages echoing similar criticisms and discouraging her from sharing about her beliefs. Even after blocking several accounts, new profiles appear and continue the behaviour. The ongoing online attention leaves Maya feeling singled out and uncomfortable, making her hesitant to post personal or cultural content.

Scenario L
Persona: Sam, 27 years old, non-binary person

Sam (he/they/them), a 27-year-old non-binary person, uses Instagram regularly to share updates about their art and travels. A few months after moving into a new apartment, they began receiving late-night messages from an anonymous account. At first, the messages seemed casual, but they quickly started referencing places Sam had recently visited and even commented on specific outfits shown in their posts. The level of detail suggested their activity was being monitored in real time. When Sam chose not to respond, the tone of the messages escalated into threats, including claims of knowing their exact address. Within days, screenshots of Sam’s profile photos, along with details about where they lived and travelled, were circulated in a public Instagram group and tagged to them directly, drawing mocking and threatening comments. The experience left Sam anxious about posting, wary of leaving home, and fearful of who might be watching.

Appendix B: Experts' Background

Table 2 Experts' background interms of years (yrs) of experience, professional background and expertise

ID	Yrs	Professional Background	Expertise
1	16	Policy or regulation, Design or user experience, Technology / engineering, Community advocacy or support services, Academia / research	Accessibility advocacy, video game accessibility research, online violence research, nonprofit policymaking
2	9	Community advocacy or support services, Academia / research	Feminist bystander behaviour projects, online harassment research, policy engagement on social media abuse
3	10	Design or user experienceTechnology / engineering, Academia / research	Responsible AI, Human Computer Interaction, online misinformation, misogynistic hate speech
4	8	Academia / research	Socio-legal research on gender, technology regulation, online misogyny, digital harms policy
5	15	Academia / research	Youth mental health, suicide prevention, online safety, digital environments and risk
6	8	Design or user experience, Technology / engineering, Academia / research	Human computer interaction research, UX research, digital wellbeing for youth
7	10	Academia / research	Information behaviour, antisocial online cultures, trolling communities, gender-based online abuse
8	15	Academia / research	Digital harms research, online abuse policy guidance, scam messaging, minority community safety
9	5	Policy or regulation	Policy advisory roles, consent education advocacy, communications and behavioural insights
10	1	Policy or regulation	Safety by Design initiative, technology-facilitated gender-based violence, platform safety policy
11	13	Design or user experience, Technology / engineering, Academia / research	UX design, social media research, policymaking, ethical design approaches

ID	Yrs	Professional Background	Expertise
12	15	Academia / research	Algorithms for bystander support, gender-based online abuse research, network polarisation
13	2	Design or user experience	UX / product design, social media user experience, digital wellbeing awareness
14	12	Design or user experience, Academia / research	Digital bias, identity misrecognition, AI harms, participatory design for marginalised communities
15	18	Policy or regulation	Cyber safety policy, Online Safety Act, internet governance, digital wellbeing policy
16	6	Design or user experience, Academia	UX design ethics, emotional harms from digital health technologies
17	25	Design or user experience, Technology / engineering Academia / research	UX research on technology-facilitated harm, lived experience perspective
18	5	Design or user experience, Technology engineering, Academia / research	Product development and scale-up, social media business models, social impact of technology
19	3	Design or user experience, Technology / engineering	Consent education co-design, gender-based violence education, product and UX design
20	6	Academia / research	Online harms against women and minorities, queer and trans experiences, deepfakes, sextortion
21	10	Policy or regulation, Design or user experience, Technology / engineering, Academia / research	Online harassment research, trust and safety governance, bias in content detection, platform regulation