Developing design strategies and policies to protect women and gender-diverse people from tech-facilitated abuse on social media

Senuri Wijenayake, Lecturer in Human-Computer Interaction (HCI)

School of Computing Technologies,

RMIT University, Australia

senuri.wijenayake@rmit.edu.au

RMIT UNIVERSITY Ready for what's next

Project Overview

Part of a larger project funded by the Australian Communications Consumer Action Network **(ACCAN)**, which aims to **understand and improve how social media safety features address tech-facilitated abuse (TFA).**

- **Stage 1 (Today): Online survey**
  Quantify how women and gender-diverse Australians engage with safety features on social media.
- **Stage 2: Co-design workshops with users**
  Develop alternative safety features based on lived experiences.
- **Stage 3: Co-design workshops with experts**
  Translate user insights into design guidelines and policy reforms.

Tech-Facilitated Abuse (TFA) on Social Media

Use of technology to harass, stalk, or intimidate individuals in unwanted, aggressive, and offensive ways.

- In the Australian context, **70%** of online harms happen in social media (e.g., abusive direct messages, coordinated harassment) [eSafety Commissioner, 2022].
- Women and gender-diverse people disproportionately targeted. Leads to severe consequences, including **withdrawal from platforms**.

Platform response:

- 30+ built-in **safety features** (tools that allow users to manage safety and privacy online) on most platforms.
- Categories: **account, post, comment, direct message, friends/follower controls.**

The gap:

- 42% of Australians report dissatisfaction with platform safety features [eSafety Commissioner, 2022].
- Only ~40% take action (mostly limited to blocking/reporting) [eSafety Commissioner, 2022].
- Limited awareness and use → **no evaluation of how women and gender-diverse people actually engage with these features.**

Our Approach

Aim: Quantify **awareness, usage, and perceived usefulness** of social media safety features against TFA.

- **Online Survey (N = 310):**
  - Conducted with women and gender-diverse Australians (30% CALD, 30% regional), focusing on **Facebook, Instagram, and TikTok**.
  - Each participant reviewed ~30 safety features on one platform.

- **Survey Structure:**
  - **Section A:** Awareness & prior use (used / aware but not used / not aware).
  - **Section B:** For 5 *used* features:
    - Purpose of use (general privacy/safety, prevent TFA, respond to TFA)
    - Perceived usefulness against strangers vs known perpetrators, for prevention and response.
  - **Section C:** For 5 *known but not used* features → reasons for non-use.

Findings: Awareness & Use of Safety Features

Low uptake: Typical users engage with **fewer than half** of available safety features, especially on **Instagram and TikTok**.

- **Most used categories** – consistent across all platforms:
    - Friends & follower controls (e.g., unfriending, blocking)
    - Account controls (e.g., private accounts)

Awareness vs use gap:

- Many users aware but not using features → Instagram (16/30), TikTok (14/28), Facebook (11/30).
- Comment controls on Instagram & TikTok show the **largest disparity** (e.g., hiding offensive comments, comment care mode).

Unfamiliarity: ~75% of users unaware of at least **1/3 of safety features**.

- Post controls (e.g., content preferences) and DM controls (e.g., safe mode, custom word filters) are least known across all platforms.

Findings: Patterns of Safety Feature Use

Routine safety focus:

- Across most feature categories and platforms, **>50% of participants** used features primarily for **general safety and privacy management**, not explicitly connecting them to tech-facilitated abuse (TFA).

TFA-related use:

- Among participants who used features in TFA-related situations, most applied them for **both prevention and response**.

Patterns by category:

- **Primarily preventative:** Account controls, post controls, comment controls on Instagram and TikTok.
- **Primarily reactive:** Comment controls on Facebook, direct message controls, friends/follower controls.

Findings: Perceived Usefulness of Safety Features

Strangers vs known individuals:

- Features are generally seen as **more effective against strangers** than known individuals.
- Gap: Features less effective for responding to TFA from people users know.

Response strategies:

- **Against strangers:** Use of default controls (who can message/follow) and assertive actions (blocking, reporting, deleting messages/comments).
- **Against known individuals:** Softer controls (muting accounts, hiding active status, safe/restricted modes) to avoid confrontation.

Findings: Perceived Usefulness of Safety Features

Prevention strategies:

- **Primarily preventative against strangers:** Comment, DM, and post controls—especially those *setting default preferences* for who can comment, message, or appear in feeds.
- **Preventing abuse from known individuals:** Certain DM controls on Facebook & Instagram, and some post controls on Instagram, perceived as more useful.

Takeaway: Existing features are adequate for prevention and response against strangers but **lack effectiveness for handling abuse from known people**.

Findings: Reasons for non use

Knowledge gaps / lack of understanding:

- Users often avoid features because they **don’t understand how they work**.
- Commonly overlooked: content controls (feed preferences, keyword filters), interaction tools (follower/limit controls), and automatically enabled features (e.g., hide offensive messages/comments).

Preference for alternatives / minimising social friction:

- Avoid features that limit interactions or could cause confrontation (blocking, reporting, private accounts, read receipts).
- Softer moderation tools (muting, snoozing) underused due to **temporary nature** and repeated effort required.

Findings: Reasons for non use

Perceived ineffectiveness / distrust:

- Users doubt that features reliably protect them or control content.
- Distrust applies to both proactive controls (content preferences, filters) and reactive tools (reporting/deleting messages/comments, blocking).

Complexity / usability challenges:

- Features that manage social connections, comment moderation, or content visibility are often seen as **too complex** (e.g., friend lists, comment filters, post audience settings).

Summary and Implications

- **Limited use:** Only a small subset of users engage extensively; the typical user uses **<50% of available features** → simply adding more features is **not sufficient**.
- **Awareness ≠ action:** Many users know about features but do not use them → need to understand barriers to use and explore ways to prompt actual engagement.
- **High unawareness:** ~75% unaware of at least a third of features → opportunity to **improve visibility, onboarding, and education** around existing tools.
- **Dual function:** Features are used **both preventatively and reactively**, highlighting their flexible role in user safety strategies.
- **Platform differences:** Clear **variances across platforms** in how features are applied against TFA, suggesting context-specific design and guidance may be needed.
- **Reasons for non-use:** Non-use stems from a combination of lack of understanding, distrust, preference for low-friction options, and perceived complexity, highlighting opportunities to simplify, educate, and build trust in safety tools.

Next steps: Co-Design Workshop

Focus: Redesign safety features identified as **particularly problematic**.

Method: Use **realistic TFA scenarios** based on literature to guide discussions.

Workshop goals:

- Observe how participants use existing features in each scenario.
- Discuss **strengths and limitations** of current tools.
- Co-develop **alternative designs** and improvements to enhance usability, effectiveness, and trust.

Hands-on, scenario-driven co-design will inform **feature redesigns that better meet user needs in real-world TFA situations**.

Thank you!

Senuri Wijenayake, Lecturer in Human-Computer Interaction (HCI)

School of Computing Technologies,

RMIT University, Australia

Contact: senuri.wijenayake@rmit.edu.au

RMIT UNIVERSITY Ready for what's next