Progress by Social Media Platforms

Twitter

Background

All four companies (Google, Meta, TikTok, and Twitter) engaged in the Tech Policy Design Labs in 2021 are working on product innovations and prototypes related to OGBV and in line with their commitments. The following changes have been made since TPDL in 2021. The progress updates listed below are based on public announcements made by Twitter.

Executive Summary

Twitter made progress both on curation (introducing Safety Mode, testing Twitter Circle and Unmentioning) and reporting (reviewing the reporting process using human-first design). Beyond those two areas, other initiatives that were highlighted as relevant to this work include: prompts to reconsider harmful tweets, publication of a safety playbook, work on hate speech lexicons. Twitter is unique among this cohort of tech companies as its open API allows entrepreneurs to build their own solutions to all manner of platform needs, including countering OGBV. Innovations in this space are available to the public beyond those produced by Twitter's own product team. However, Twitter's progress is somewhat overshadowed by the potential risks associated by the recent change in leadership.

Curation

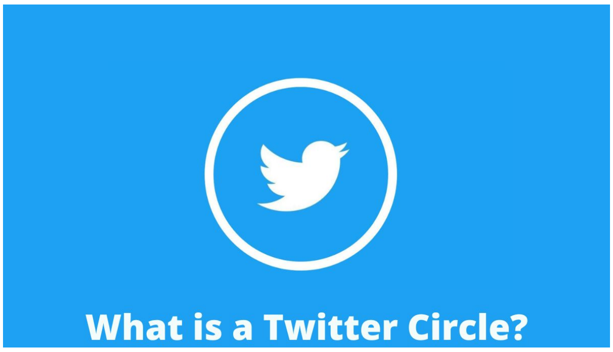
Twitter has made substantial progress against the commitments in both areas. On curation, it has announced the following features – now all available globally:

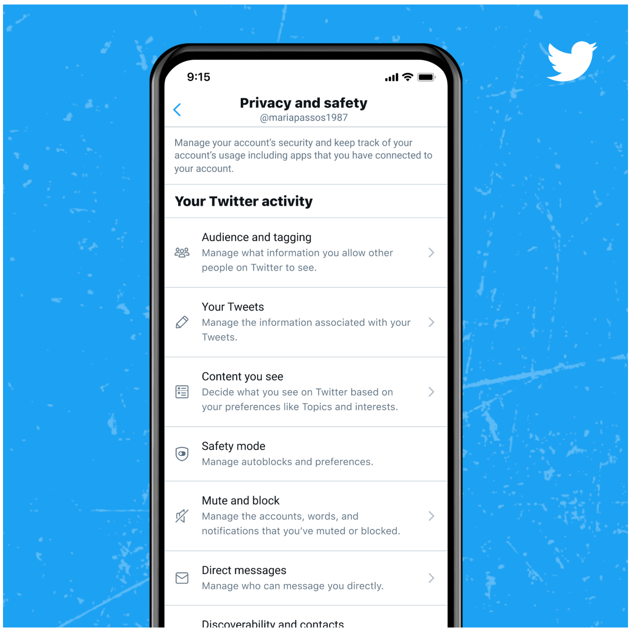
- **Safety Mode**: temporarily blocks accounts for seven days for using potentially harmful language – such as insults or hateful remarks – or sending repetitive and uninvited replies or mentions (optional setting). Twitter is currently testing it directly with users and tracking adoption rates to understand its potential impact.

- **Twitter Circle:** lets users add up to 150 people who can see their Tweets when they want to share “with a smaller crowd.” (It can be noted that Instagram had introduced a similar feature in 2018 (“Close friends”)

- **Unmentioning:** allows users to remove themselves from conversations they don’t want to be a part of

- **Mid-Conversation Control**: Twitter added the possibility to change the settings of who can reply to their tweets midway through a conversation




9:15

Privacy and safety
@mariapassos1987

Manage your account’s security and keep track of your account’s usage including apps that you have connected to your account.

Your Twitter activity

Audience and tagging
Manage what information you allow other people on Twitter to see.

Your Tweets
Manage the information associated with your Tweets.

Content you see
Decide what you see on Twitter based on your preferences like Topics and interests.

Safety mode
Manage autoblocks and preferences.

Mute and block
Manage the accounts, words, and notifications that you’ve muted or blocked.

Direct messages
Manage who can message you directly.

Discoverability and contacts

Reporting

Twitter started testing in Dec 2021 an overhaul reporting process aiming to make it easier for people to alert them of harmful behavior. This new Report Tweet flow is now available globally (Twitter communicated it was available “in all languages” – further research is needed to clarify this.). Building on human-centered design, this new approach lifts the burden from the individual to be the one who has to interpret the violation at hand. Instead it asks them what happened. This method is called symptoms-first. Twitter communicated it has enabled an increase of 50% of actionable reports during testing.

Twitter’s updated policy in December 2021 was seen as an important change to previous reporting mechanisms by the sector which did not previously centre the experience of the victim/survivor.

Other

Beyond their specific TPDL commitments around curation and reporting, one should note the following work done by Twitter since mid 2021:

- **Keep experimenting prompts** to users to reconsider harmful tweets. This approach aligns with demand from civil society for platforms to intervene prior to offensive content being posted. Twitter communicated about "a shift towards more proactive strategies". The prompt has now been developed in English, Portuguese, Spanish and Arabic. Peer reviewed research published in May 2022 concluded this feature led to 6% fewer offensive Tweets (based on the analysis of 200,000 tweets). This study also suggests that people who are exposed to a prompt are slightly less likely to compose future offensive replies.

- **Communication on safety tools**: Twitter created and published an updated safety playbook and have launched various promoted tweet campaigns featuring existing and new safety tools-related videos. They found innovative ways to educate people via partnerships to highlight recent features like removing followers, Safety Mode, and conversation settings.

- **Open API**: Twitter is also unique among this cohort of tech companies with its open API that allows researchers to access data more easily, and entrepreneurs to build their own tools working on the platform including countering OGBV. Innovations in this space are available to the public beyond those produced by Twitter's own product team.

- **Work on hate speech lexicons**: Twitter was noted for its approach to hate speech lexicons. Which directly links to protected characteristics and therefore people who are marginalized and more vulnerable to online gender based violence including "promot[ing] violence against or directly attack[ing] or threaten[ing] other people on the basis of race, ethnicity, national origin, caste, sexual orientation, gender, gender identity, religious affiliation, age, disability, or serious disease. We also do not allow accounts whose primary purpose is inciting harm towards others on the basis of these categories."

- **Live Location Sharing:** Twitter has updated their Privacy Information Policy to prohibit sharing someone else's live location in the majority of cases, due to the increased risk of physical harm. Going forward, they will remove tweets that share live location information, and accounts that are targeted towards sharing individuals' live location will be suspended.

However, Elon Musk's acquisition of the platform in 2022 has put safety at risk. Among other measures like cutting Twitter workforce by half**, he disbanded the Trust and Safety Council**, a volunteer group that was established in 2016 to advise the platform on site decisions. Instead, Musk has announced he will form a Content Moderation Council. Although this Council was perceived as very important to civil society experts, they questioned whether the format of relatively short calls with Trust and Safety Council members constituted meaningful engagement with CSOs, particularly where OGBV is not a primary focus of discussion in these forums. Some Global South members of the Trust and Safety Council explicitly expressed frustration that they are not engaged as co-creators but are presented with policy recommendations after Twitter has produced them and are asked for advice, with no accountability on whether their input is integrated.