Tackling Online Abuse: Written evidence submitted by Twitter on 13/09/2021 (TOA0012)

Summary

We are pleased to provide a written submission to the Petitions Committee call for evidence to support the inquiry into tackling online abuse. This follows our engagement with the 2018 inquiry into online abuse and the experiences of disabled people. Our response also builds on the written and oral evidence we have provided to support the development of the UK Government's Online Harms White Paper.

We are pleased to include information about our approach to online abuse, alongside sections on how we develop our policies, our approach to enforcement and detail about our rules.

Twitter’s approach to online abuse

In order to facilitate healthy dialogue on the service, and empower individuals to express diverse opinions and beliefs, we prohibit behaviour that harasses or intimidates others. In addition to posing risks to people’s safety, abusive behaviour may also lead to physical and emotional hardship for those affected.

In 2018, we shared that building a Twitter free of abuse, spam, and behaviour that distracts from the public conversation is our top priority. Since then, we have made strides in creating a healthier service. There will always be more to do, but we have made meaningful progress that is important to share:

- In October 2019, we announced that more than 50% of abusive content that is enforced is now surfaced proactively to our teams for review, instead of relying on reports from people on Twitter.
- We also announced a 105% increase in accounts actioned by Twitter (locked or suspended for violating the Twitter Rules); concurrently, we are now suspending 3 times more abusive accounts within 24 hours after a report compared to the same time last year.
- In June 2019, we refreshed the Twitter Rules with simple, clear language and reorganised them into high-level categories: safety, privacy and authenticity. Over the years we added new rules and updated existing ones, but these changes eventually made our rules confusing and difficult to understand. With this change, we have also gone from approximately 2,500 words to under 600.

Transparency is a key guiding principle in our mission to serve the public conversation. Since 2012, our biannual Twitter Transparency Report has highlighted trends in requests made to Twitter from around the globe, and now includes enforcement statistics regarding the Twitter Rules.

In terms of Twitter’s proactive approach to dealing with issues such as abuse, hateful conduct, and other malicious activities, we increasingly look for behavioural signals, not just content, as a way to identify content that may violate our rules. On abuse, we are tackling issues of behaviours that distort and detract from the public conversation by integrating behavioural signals into how Tweets are presented. There are many signals we are taking in, most of which are not visible externally. Examples include if an account has not confirmed their email address; if the same person signs up for multiple accounts simultaneously; or accounts that repeatedly Tweet and mention accounts that do not follow them.

These signals are now considered in how we organise and present content in communal areas like Conversations and Search, leading to reductions in the number of abuse reports in both.

We are acutely aware that many high profile users can, at times, be particularly vulnerable to abuse and harassment. In February 2020, we created a series of videos with high-profile influencers in the UK to talk about the experiences they have had and safety tools they have used. Garnering 1.8 million impressions - and with an average view through rate of 57% - we saw that this was a way to use engaging and authentic content to reach a wider range of users on the service. We have also deepened our collaborations focused on specific types of users who may be vulnerable to abuse - since the beginning of 2019, we have partnered with the Parliamentary Security Department to rapidly identify and remove Tweets breaking our rules that are targeting Members of Parliament. By providing an expedited reporting channel to Twitter, and meeting once a week to discuss any ongoing incidents, this ensures we can respond quickly to MPs being targeted with abuse.

More recently, we have also introduced a number of product changes and experiments. Critically, on 11 August 2020, we made conversation controls available to all users following a trial in the spring. Before you Tweet, you can now choose who can reply with three options: 1) everyone (standard Twitter, and the default setting), 2) only people you follow, or 3) only people you mention. Tweets with the latter two settings will be labeled and the reply icon will be grayed out for people who can't reply. Our trial identified that people who face abuse find these settings helpful - those who have submitted abuse reports are three times more likely to use these settings.

In 2020, we began testing prompts that encourage people to pause and reconsider a potentially harmful or offensive reply — such as insults, strong language, or hateful remarks — before Tweeting it. Once prompted, people had an opportunity to take a moment and make edits, delete, or send the reply as is.

These tests ultimately resulted in people sending less potentially offensive replies across the service, and improved behavior on Twitter. We learned that:

- If prompted, 34% of people revised their initial reply or decided to not send their reply at all.
- After being prompted once, people composed, on average, 11% fewer offensive replies in the future.
- If prompted, people were less likely to receive offensive and harmful replies back.

We'll continue to explore how prompts — such as reply prompts and article prompts — and other forms of intervention can encourage healthier conversations on Twitter. Our teams will also collect feedback from people on Twitter who have received reply prompts as we expand this feature to other languages.

Furthermore, in September 2021, we introduced Safety Mode, a new feature that aims to reduce disruptive interactions. We are rolling out this safety feature to a small feedback group, beginning with accounts that have English-language settings enabled.

Safety Mode is a feature that 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. When the feature is turned on in your Settings, our systems will assess the likelihood of a negative engagement by considering both the Tweet's content and the relationship between the Tweet author and replier. Our technology takes existing relationships into account, so accounts you follow or frequently interact with will not be autoblocked.

We will continue to observe how Safety Mode is working as part of our work to empower people with the tools they need to feel more comfortable participating in the public conversation.

Our approach to policy development and enforcement philosophy

Twitter is reflective of real conversations happening in the world and that sometimes includes perspectives that may be offensive or controversial to others. While we welcome everyone to express themselves on our service, we will not tolerate behaviour that harasses, threatens, or uses fear to silence the voices of others.

As outlined, our Twitter Rules are in place to help ensure everyone feels safe expressing their beliefs and we strive to enforce them with uniform consistency. For more information, read about our different enforcement actions.

Our policy development process

Creating a new policy or making a policy change requires in-depth research around trends in online behaviour, developing clear external language that sets expectations around what is allowed, and creating enforcement guidance for reviewers that can be scaled across millions of Tweets.

While developing our policies, we gather feedback from a variety of internal teams as well as our Trust & Safety Council. In 2016, we established the Twitter Trust and Safety Council, which brings together more than 40 experts and organisations (including in the UK) to help advise us as we develop our products, programs and policies. In 2019, we broadened and restructured the Council, to include a more diverse range of voices and organise members to have deeper conversations.

We have also started to consult publicly on new policies. In 2018, we ran a public consultation around new rules on dehumanising language - in two weeks, we received more than 8,000 responses from people located in more than 30 countries. Building on this in 2019, we ran a public survey on our initial draft of new rules around synthetic and manipulated media, gathering more than 6,500 responses from people around the world. The results of the survey are available here. Most recently, in March this year we called for responses to a public survey to help inform the development of our policy framework for world leaders. Nearly 49,000 people from around the globe took time to share their feedback on how content from world leaders should be handled on our service.

Engagement like this is vital to ensure we are considering global perspectives around the changing nature of online speech, including how our rules are applied and interpreted in different cultural and social contexts.

Finally, having developed our policy, we train our global review teams, update the Twitter Rules, and start enforcing.

Our enforcement philosophy

We empower people to understand different sides of an issue and encourage dissenting opinions and viewpoints to be discussed openly. This approach allows many forms of speech to exist on our service and, in particular, promotes counter-speech: speech that presents facts to correct misstatements or misperceptions, points out hypocrisy or contradictions, warns of offline or online consequences, denounces hateful or dangerous speech, or helps change minds and disarm.

Thus, **context matters**. When determining whether to take enforcement action, we may consider a number of factors, including (but not limited to) whether:

- the behaviour is directed at an individual, group, or protected category of people;
- the report has been filed by the target of the abuse or a bystander;
- the user has a history of violating our policies;
- the severity of the violation;
- the content may be a topic of legitimate public interest.

Learn more about our enforcement philosophy.

Tools and reporting

There are a number of tools available to our users to manage their experience:

- If you see or receive a reply you don't like, unfollow the account.
- If the behaviour continues, we recommend that you block the account. Blocking will prevent that person from following you, seeing your profile image on their profile page, or in their timeline; additionally, their replies or mentions will not show in your Notifications tab (although these Tweets may still appear in search).
- Use conversation controls to choose who can reply to your Tweets.

If you are receiving unwanted, targeted and continuous replies on Twitter and feel it constitutes online abuse, we encourage our users to consider reporting the behaviour to Twitter. For Twitter, multiple reporting mechanisms are made available in-app or via our website (where there is also a step-by-step guide). To better support users who become victims of certain behaviours, our reporting process has been refined and simplified in recent years, reducing the number of ‘clicks’ required to submit a report by more than 50%.

We encourage our users to take threats seriously. As stated on our website, if a user believes that they are in physical danger, we advise them to contact local law enforcement authorities. We also provide the following advice for if an individual does decide to work with law enforcement:

- Document the violent or abusive messages with print-outs or screenshots
- Be as specific as possible about why you are concerned
- Provide any context you have around who you believe might be involved, such as evidence of abusive behaviour found on other websites
- Provide any information regarding previous threats you may have received

In addition, we encourage our users to reach out to the people they trust. When dealing with negative or harmful interactions, it can help to turn to family and friends for support and advice. Our website contains suggestions for how to help a friend or family member with online abuse. There are also many online safety resources that can help.

Twitter Rules and policies to protect against online abuse

Twitter Rules are in place to help ensure everyone feels safe expressing their beliefs and we strive to enforce them with uniform consistency. Critically, we prohibit:

- **Abusive behaviour:** You may not engage in the targeted harassment of someone, or incite other people to do so. We consider abusive behaviour an attempt to harass, intimidate, or silence someone else's voice.

- **Abusive profile information:** You may not use your username, display name, or profile bio to engage in abusive behaviour, such as targeted harassment or expressing hate towards a person, group, or protected category.

- **Hateful conduct:** You may not promote violence against or directly attack or threaten 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.

- **Hateful imagery and display names:** You may not use hateful images or symbols in your profile image or profile header. You also may not use your username, display name, or profile bio to engage in abusive behaviour, such as targeted harassment or expressing hate towards a person, group, or protected category.

Some of the types of behaviour we will also take action on include:

- **Wishing or hoping serious harm on a person or group of people:** We do not tolerate content that wishes, hopes, promotes, incites, or expresses a desire for death, serious bodily harm or serious disease against an individual or group of people.

- **Unwanted sexual advances:** We prohibit unwanted sexual advances and content that sexually objectifies an individual without their consent.
- **Using aggressive insults with the purpose of harassing or intimidating others:** We take action against excessively aggressive insults that target an individual, including content that contains slurs or similar language.
- **Encouraging or calling for others to harass an individual or group of people**: We prohibit behaviour that encourages others to harass or target specific individuals or groups with abusive behaviour. This includes, but is not limited to; calls to target people with abuse or harassment online and behaviour that urges offline action such as physical harassment.

Our Rules are available in full here. On our website, we provide a number of examples of Tweets that would violate our rules (for instance, information about our Hateful Conduct and Abuse policies).

An individual does not need to be the target of abusive content or behaviour for it to be reviewed for violating the Twitter Rules. We review both first-person and bystander reports of such content. To help our teams understand the context of a conversation, we may need to hear directly from the person being targeted, to ensure that we have the information needed prior to taking any enforcement action.

For examples of abusive behaviour which violates the Twitter Rules, please see here.

We are grateful to the Petitions Committee for giving us the opportunity to provide detailed feedback on our approach to tackling online abuse, and we look forward to continuing to work with the Committee and government on these issues.

Twitter's purpose is to serve the public conversation, and we continue to strive to facilitate healthy dialogue on the service, empower individuals to express diverse opinions and beliefs, and prohibit behaviour that harasses or intimidates, or is otherwise intended to shame or degrade others.