Twitter UK
20 Air Street
London
W1B 5AN

twitter.com

5 September 2018

Dear Lord Bew,

Thank you for your letter (16 August 2018). I am writing to update you on the work we have been doing at Twitter.

We have committed Twitter to help increase the collective health, openness, and civility of public conversation, and to hold ourselves publicly accountable towards progress. We believe we must commit to a rigorous and independently vetted set of metrics to measure the health of public conversation on Twitter. We have committed to sharing our results publicly to benefit all who serve the public conversation.

To try to achieve this goal, we have made a number of changes, including enhanced safety policies, better tools and resources for detecting and stopping malicious activity, tighter advertising standards, and increased transparency to promote public understanding of all of these areas.

In furtherance of our desire to improve accountability and explore potential ways of measuring conversational health, earlier this year we issued a request for proposals (RFP) to collaborate with the world’s leading experts in this domain. We received more than 230 proposals from global institutions eager to help us further examine a broad range of topics that will advance this important work. We announced in August that we had selected two partners; the first team, led by scholars from Leiden University, will look at how echo chambers form and their effect, as well as the difference between incivility and intolerance within Twitter conversations. The second research project will be led by Professor Miles Hewstone and John Gallacher at The University of Oxford, in partnership with Marc Heerdink at the University of Amsterdam. They will be examining how people use Twitter, and how exposure to a variety of perspectives and backgrounds can decrease prejudice and discrimination.

On a practical level, since the beginning of 2017, we have introduced over 30 product and policy changes. These changes aim to ensure that everyone on Twitter is free to express themselves and to feel safe in doing so. We have launched our Safety Center and continue to work with leading safety experts from around the world through our Trust & Safety Council, while seeking broader input from a range of organisations. In the UK, for example, the organisations we have engaged with on these issues include the Community Security Trust, Institute for Strategic Dialogue and Media Diversity Institute. We do not tolerate behaviour that harasses, intimidates, or uses fear to silence another person’s voice; our Hateful Conduct Policy can be viewed in full on our Safety Center. We continue to not only launch new policies, but also update existing ones on an ongoing basis to reflect the changing nature of activity on our platform, such as dehumanisation and harassment.

Technology has huge potential in this space, but it is not currently possible to automate the review and removal of content using technology without extremely high numbers of false positives. Natural language processing is an important tool but, it is not yet advanced enough to detect intimidatory content - itself a subjective standard - without wrongly catching speech such as sarcasm, satire, legitimate criticism and consensual conversations where the context changes the nature of the speech.

To further put the challenge into context, fewer than 1% of accounts make up the majority of accounts reported for abuse, but a lot of what is reported does not violate our rules. While still a small overall number, these accounts have a disproportionately large – and negative – impact on people's experience on Twitter. The challenge for us has been: how can we proactively address these disruptive behaviours that do not violate our policies but do negatively impact the health of the conversation?

Today, we use policies, human review processes, and machine learning to help us determine how Tweets are organised and presented in communal places, like conversations and search. Now, we are tackling issues of behaviours that distort and detract from the public conversation in those areas by integrating new signals into how Tweets are presented. By using new tools to address this conduct from a behavioural perspective, we are able to improve the health of the conversation, and everyone's experience on Twitter, without waiting for people who use Twitter to report potential issues to us.

There are many new signals we are taking in, most of which are not visible externally. Just a few examples include: if an account has not confirmed their email address, if the same person signs up for multiple accounts simultaneously, accounts that repeatedly Tweet and mention accounts that do not follow them, or behaviour that might indicate a coordinated attack. We are also looking at how accounts are connected to those that violate our rules, and how they interact with each other. Early testing saw a positive impact with a 4% drop in abuse reports from search and 8% fewer abuse reports from conversations.

We have also expanded our use of technology to improve the speed we review and action reports from users. We are now taking action on 10 times the number of abusive accounts every day compared to the same time last year. We limit account functionality, or place suspensions, on thousands more abusive accounts each day. Indeed, accounts that demonstrate abusive behaviour may now be limited for a time, and told why. Those accounts put into this period of limited functionality generate 25% fewer abuse reports - and approximately 65% of these accounts are in this state just once. Earlier this year, our new systems were able to remove twice the number of repeat offenders who create new accounts after being suspended for violations. We have also invested in our teams and resources by acquiring Smyte, a technology company that specialises in safety, spam and security issues.

To better support users who become victims of this behaviour, Twitter's reporting process has been refined and simplified in recent years, reducing the number of ‘clicks’ required by more than 50%. We have also introduced new ways for users to manage their experience, including notification filters and muting keywords. For example:

- Our Quality Filter works to identify potentially low-quality (but not necessarily abusive) activity. The Filter has led to fewer unwanted interactions: we have seen a 40% reduction in the number of people blocking other users after another user they do not follow tries to engage with them.
- We have given users control over what they see in search results through Safe Search mode. These filters exclude potentially sensitive content, along with accounts the user has muted or blocked, from search results. Users have the option to turn it off, or back on, at any time (instructions outlined below).
- Twitter now also uses technology to identify and support users seeing an increase in activity. Now, we enable a full screen prompt to the user to encourage them to review their notifications filters. Users can be taken to their settings to change their notifications; our hope is that those who are seeing an uptick in engagement (because of a viral Tweet or something more malicious) can limit what can be an overwhelming experience.

We have a range of enforcement actions against accounts posting abusive content, which can be taken at the Tweet, Direct Message, and Account levels - or sometimes a combination of the three. At the Tweet level, we now hide reported Tweets while a decision is being made, at which point we could limit the visibility of the Tweet or require the user to delete it. For Direct Messages, we may place a violating Direct Message behind an interstitial to ensure no one else in the group can see it again; or simply stop conversations between a reported violator and the reporter’s account. And at the Account level, actions we may take include requiring media or profile edits, placing an account in read-only mode, verifying account ownership and permanent suspension. There are actions we may take even when a Tweet does not violate our policies, but it may be necessary to limit its visibility. A full breakdown of the range of enforcement actions available is provided in **Appendix I**.

When online abuse escalates, legal requests may come from law enforcement, government agencies, lawyers representing a criminal defendant, civil litigants, or from other authorised reporters. Twitter has developed detailed guidelines for law enforcement to support their work, and launched an online legal submissions site to make the process of reporting to Twitter more efficient. We regularly host training and engagement sessions with law enforcement, in the UK and abroad, to ensure our processes and tools are understood by relevant officers. And our proprietary technology continues to disrupt the use of Twitter for the promotion of terrorism. In the latest reported period, for instance, a total of 274,460 Twitter accounts were permanently suspended for violations related to promotion of terrorism between July 2017 and December 2017. Of those suspensions, 93% consisted of accounts flagged by our internal, proprietary spam-fighting tools, while 74% of those accounts were suspended before their first tweet.

For those in Parliament specifically, we are making our Partner Support Portal available to the relevant security teams, enabling our trusted partners to expedite tickets for MPs and Lords. We continue to have strong relationships with different teams across Parliament and are committed to improving the support we provide to them. This includes newly developed tools to monitor for attempts to manipulate election conversation. We are also committed to supporting a dedicated process during elections to provide support to candidates.

We are continuously building on our products to give more transparency and information to our users. We started this year with our latest Twitter Transparency Report, adding a new section to our legal removals report covering requests to remove content from verified journalists and other media and news outlets. We will continue to build on our Transparency Report on an ongoing basis.

Twitter stands for freedom of expression and people being able to see all sides of any topic - that is put in jeopardy when abuse and harassment stifle and silence those voices. While there is still much work to be done, we are making progress.

I would be very happy to meet in person and talk through any of the points above in greater detail. Our policies and enforcement options evolve continuously to address emerging behaviours online. We will continue to expand and update them to respond to the changing contours of online conversation.

Thank you for reading.

Katy Minshall
Head of UK Government, Public Policy and Philanthropy

## Appendix I: Enforcement Actions

Our enforcement actions against accounts posting abusive content can be taken at the Tweet, Direct Message, and accounts levels, and sometimes employ a combination of these.

*Tweet-level enforcement*
We take action at the Tweet level to ensure that we are not being overly harsh with an otherwise healthy account that made a mistake and violated our Rules. A few of the ways in which we might take action at the Tweet level include:

- Limiting Tweet visibility: This makes content less visible on Twitter, in search results, replies, and on timelines. Limiting Tweet visibility depends on a number of signals about the nature of the interaction and the quality of the content.
- Requiring Tweet deletion: When we determine that a Tweet violated the Twitter Rules, we require the violator to delete it before they can Tweet again. We send an email notification to the violator identifying the Tweet(s) in violation and which policies have been violated. They will then need to go through the process of deleting the violating Tweet or appealing our review if they believe we made an error.
- Hiding a violating Tweet while awaiting its deletion: In the interim period between when Twitter takes enforcement action and the person deletes the Tweet, we hide that Tweet from public view.

*Direct Message-level enforcement*

- Stopping conversations between a reported violator and the reporter's account: In a private Direct Message conversation, when a participant reports the other person, we will stop the violator from sending messages to the person who reported them. The conversation will also be removed from the reporter's inbox. However, if the reporter decides to continue to send Direct Messages to the violator, the conversation will resume.
- Placing a Direct Message behind an interstitial: In a group Direct Message conversation, the violating Direct Message may be placed behind an interstitial to ensure no one else in the group can see it again.

*Account-level enforcement*
We take action at the account level if we determine that a person has violated the Twitter Rules in a particularly egregious way, or has repeatedly violated them even after receiving notifications from us.

- Requiring media or profile edits: If an account's profile or media content is not compliant with our policies, we may make it temporarily unavailable and require that the violator edit the media or information in their profile to come into compliance. We also explain which policy their profile or media content has violated.
- Placing an account in read-only mode: If it seems like an otherwise healthy account is in the middle of an abusive episode, we might tempo-

rarily make their account read-only, limiting their ability to Tweet, Retweet, or Like content until calmer heads prevail. The person can read their timelines and will only be able to send Direct Messages to their followers. When an account is in read-only mode, others will still be able to see and engage with the account. The duration of this enforcement action can range from 12 hours to 7 days, depending on the nature of the violation.

- Verifying account ownership: To ensure that violators do not abuse the anonymity we offer and harass others on the platform, we may require the account owner to verify ownership with a phone number or email address. This also helps us identify violators who are operating multiple accounts for abusive purposes and take action on such accounts. Note that when an account has been locked pending completion of a challenge (such as being required to provide a phone number), it is removed from follower counts, Retweets, and likes until it provides a phone number.
- Permanent suspension: This is our most severe enforcement action. Permanently suspending an account will remove it from global view, and the violator will not be allowed to create new accounts. When we permanently suspend an account, we notify people that they have been suspended for abuse violations, and explain which policy or policies they have violated and which content was in violation.

Violators can appeal permanent suspensions if they believe we made an error. They can do this through the platform interface or by filing a report. Upon appeal, if we find that a suspension is valid, we respond to the appeal with information on the policy that the account has violated.

*Actions we may take against non-violating content*

- Placing a Tweet behind an interstitial: We may place some forms of sensitive media like adult content or graphic violence behind an interstitial advising viewers to be aware that they will see sensitive media if they click through. This allows us to identify potentially sensitive content that some people may not wish to see. Learn more about how to control whether you see sensitive media.
- Withholding a Tweet or account in a country: We may withhold access to certain content in a particular country if we receive a valid and properly scoped request from an authorized entity in that country. We also clearly indicate within the product when content has been withheld. Read more about country withheld content.