Twitter UK
20 Air Street
London
W1B 5AN

twitter.com

16 December 2021

Dear Chair,

Thank you for your letter. Please see below my response to your questions.

Rules and enforcement: Mr Sunder Katwala's Twitter thread

Our policy is not to review hypothetical Tweets, as context is critical - for example, documenting an example of racist abuse that has been received on another online service. This means that a specific sentence might be prohibited in one conversation, for instance, but permitted in another context. However, I fully appreciate the Committee's concerns and can confirm Tweets like these, were they real, would be reviewable under our Hateful Conduct policy, which prohibits abuse that targets individuals or groups of people belonging to protected categories. 2 of the 3 Tweets the Committee shared from Mr Katwala's thread would almost certainly have already broken our rules, were they real - however, as my colleague Nick Pickles also highlighted at a separate Parliamentary hearing in October:

"*We are planning on rolling out a stronger policy specifically addressing people who say that someone cannot be from a certain country because of the colour of their skin. That policy gap is one that we are strengthening. That was identified during the Euros as a particular issue where we were not taking action, but certain types of abuse were targeted using that language."*

This will be completed early in the new year to clearly address all circumstances of the type of harmful abuse Mr Sunder Katwala's Twitter thread highlights, and I will share the full details with the Committee.

On the additional Tweet shared in your most recent letter, which featured a screenshot of several accounts, all but one were no longer on the service already; only one account was still live, which has now received a temporary suspension.

Rules and enforcement: reporting of racist abuse

Racist abuse is not permitted on Twitter. Our Hateful Conduct Policy and our Abusive Behaviour Policy provide details and examples of the behaviours that we would consider in breach of our

rules, covering a range of types of Tweets that could constitute racist abuse. I can confirm that *"you may not promote violence against or directly attack or threaten other people"* is not discretionary, and it is prohibited in all circumstances. We completely agree that racist abuse can consist of certain emojis and hateful imagery, and already view them as violative beyond our specific rules on the incitement of fear or the incitement of others to harass.

The global review of our Hateful Conduct policy is ongoing; the terms of reference are to assess how effectively our current policy is addressing harmful conversations on the service, and directly addresses the concerns raised in the above section 'Mr Sunder Katwala's Tweet thread.' In broad terms, our policies are reviewed and updated on an ongoing basis by our Trust and Safety Teams and by our Trust and Safety Council to address new vectors of abuse. As a recent example, on Monday (13th December) we announced that our Hateful Conduct dehumanisation policy now covers all protected categories. This means that we prohibit language that dehumanises others on the basis of religion, caste, age, disability, disease, race, ethnicity, national origin, gender, gender identity, or sexual orientation.

Since I appeared before the Committee, we have also commenced testing of a new reporting flow to make it easier for people to alert us of harmful behavior. Reporting harmful Tweets should be straightforward and lift the burden from the individual to be the one who has to interpret the violation at hand. Further information is available here.

Signify research

Alongside colleagues from our enforcement team and Twitter Developer Platform, I met with Signify in October. We discussed the discrepancies between their data and ours; we understand that Signify are using a Football Association rule to define discriminatory abuse, which lays out a definition of what a footballer would be suspended for. We discussed that what someone may be suspended for in a professional context will not always match with the Twitter rules for speech that is and is not permitted on our service. However, Signify shared that in re-reviewing their data almost all of the racist abuse they had identified had already been removed.

The changes I referred to in my letter are improvements to our proactive detection over the 2020-21 season - in effect, more effective technology we developed to detect this abuse. This meant that for Euro 2020, in total, over 90% of the Tweets we removed for abuse over this period were detected proactively. This season (since August 7), we have removed 97% of Tweets targeting the football conversation with violations of our rules which without requiring reports.

During our meeting, we were pleased that Signify shared that they were keen to work collaboratively - we have invited them to meet again in the new year. We also continue to meet with the PFA and other football authorities on a very regular basis to share the work we are doing and work collectively to address online racist abuse.

Finally, I can confirm we have reviewed all Tweets provided to us by Signify (and clarify that the spreadsheet provided to us did indeed have 1781 rows; not 1772).

Please do let us know if you have any further questions.

Yours sincerely,

Katy Minshall
Head of UK Public Policy