10 Brock Street
London
NW1 3FG

Thursday 16th December, 2021

Dame Diana Johnson MP,
Home Affairs Committee,
House of Commons,
London, SW1A 0AA.

Dear Dame Diana Johnson MP,

Thank you for your Committee's letter on 1st December 2021.

Moderation and reporting processes

Your letter set out a number of questions about our policies and their enforcement, with particular regard to the action we take on emojis.

No one should have to put up with racist abuse anywhere, and we don't want it on Instagram. Our policies do not allow hate speech in written or visual form, which includes emojis, as outlined in our hate speech policy here. We also know that we have a responsibility to look at how language evolves, which includes emojis and other forms of visual communication.

In relation to the horrendous abuse we saw on the accounts of the England team players following the Euros, we removed 15,000 comments from 11-13th July for Bullying and Harassment and Hate Speech directed at members of the England team, the majority of which were found and flagged proactively using our AI technology. The majority of these comments were also removed within an hour of them being posted, and the vast majority within 24 hours. However, the Committee subsequently raised 5 pieces of content on 8th September, all of which we removed within minutes of investigating, as well as disabling several accounts for repeatedly violating our policies.

Your letter asked about why content that may violate policies could be on Instagram and what has changed since July. Unfortunately zero tolerance does not mean zero incidence. As

discussed at the hearing and in my earlier letter to you, AI detection and indeed human review are not perfect and we are sorry to see these comments on Instagram a few weeks after being posted. When it comes to hate speech, context is extremely important. Words, emojis and phrases, on their own, may not be inherently violating - often we need to see the context in which they are being used, as well as who is sending them and to whom, and AI alone does not always get this right. We also need to understand whether the words or emojis that may be used to abuse someone, are being used to condemn or raise awareness of hate speech, and whether certain words that have been used to racially abuse are being used as reclaimed speech. AI technology and human review are therefore both important, but so too is in-app reporting.

While we swiftly removed the 15,000 comments referenced above, after the Euros final there were some instances where we were incorrectly telling people that the use of emojis in a hateful context did not violate our policies and this was a mistake. We want to be clear: using emojis to abuse people based on their race absolutely violates our hate speech policy. We use AI technology to try and prioritise reports for our reviewers - looking at factors such as virality, severity of harm and likelihood of violation. Following the Euros final we found that our AI had been mistakenly marking some comments containing emojis attacking players as benign, when they were not. Since then we have made several changes to work to help ensure people are not seeing comments like this on Instagram, including adding certain strings of emojis that are nearly always used in an abusive context to our automatic filters. We encourage our community and expert partners to continue to flag content where they believe we made the wrong decision and to use our in-app appeal processes.

I want to be clear though that, so long as people are attempting to send hate to professional athletes online and constantly adapting their language or methods of communication in order to do so, we will not see zero incidence. In this context, it is vital that we not only constantly improve and update our policies, AI enforcement and human review processes, but that we also find ways to stop hate like this appearing in the first place, reduce its distribution when it does and work with others to address the causes of this hatred so it does not reappear elsewhere - online or offline.

Some of the other steps we have taken this year were set out in my previous letter and include: stricter penalties for those who send abusive messages; an enhanced comment warning that we estimate causes around 50% of attempted, hateful comments to be changed or deleted in a given week; a feature called Hidden Words that, when turned on, filters comments and DMs containing abusive words, phrases or emojis so people never have to see them; and a Limits feature which lets people hide comments and DMs from people who don't follow them or only recently followed them, as we know most abuse comes from these groups. We are working directly with football bodies to support players in turning these tools on, to help protect their accounts.

In addition, it's important that we continue to work with others when it comes to societal issues like racism. Where online behaviour may also be illegal, we also respond quickly to

valid legal requests which can help police investigations and during the Euros the NPCC Football Lead called our cooperation 'excellent'. We are in ongoing discussions with the National Police Chiefs Council, the UK Football Policing Unit and relevant local police forces to understand how we can continue to best support active investigations and ensure valid data requests are successful in accordance with applicable law and our terms of service. We are also proud to be part of the Football Against Online Hate working group alongside representatives from the Premier League, FA, PFA, EFL, Women in Football, Kick it Out, Sky and others across the industry, to collectively tackle this issue.

Your letter asked about transparency and since we last wrote to you we published our most recent Community Standards Enforcement report. This report showed that in July to September this year - including the Euros final and its aftermath - we took action on 6 million pieces of hate speech content globally, of which 93.8% was found before anyone reported it. This is a significant increase over recent years, but also means that just over 6% of hate speech content we removed in that quarter was reported to us before it was flagged by our technology.

For the first time, we also included in our report an estimate of the prevalence of hate speech on Instagram. This showed that, from July to September this year, we estimate that for 10,000 views of content on Instagram, only about 2 would contain hate speech. This report also contained detail on how we are addressing prevalence - beyond just removing content - and this month we also published detail on new AI technology we have developed to further improve accurate detection and deletion of violating content moving forward.

Finally, with regards to our policies themselves, we regularly pressure test our approach with safety experts and policymakers and make changes as needed, including since the summer. For example, in October, and after extensive consultation, we have updated our bullying and harassment policies in order to give more protections for public figures, including athletes. As part of this process we consulted our expert advisors, politicians, journalists, representatives of different communities with lived experience, content creators and public figures, and we will continue to work with experts and listen to members of our community to ensure our platforms remain safe.

Meta's approach to hate speech

I would disagree with the assertion that our company does not take sufficient, swift and wide-scale action across our platforms regarding hateful content. As set out above we take wide scale and prompt action to address violations of our policies - and to try and discourage them from happening in the first place - and this includes hateful content such as racist abuse.

People do not want to see hate and abuse when they use our apps and our advertisers don't want their ads next to it either. That's why we have clear rules about what isn't allowed on our platforms, are on track to spend more this year on safety and security than any other

tech company, even adjusted for scale, and have over 40,000 people to keep people safe on our apps. We have spent more than $13 billion on safety and security, and almost halved the amount of hate speech people see on Facebook over the last three quarters, and on Instagram we estimate only 2 in 10,000 views of content may be hate speech.

Helping people have positive experiences on our apps by seeing content that is more relevant and valuable to them is in the best interests of our business over the long term. Our systems are not designed to reward provocative content. In fact, key parts of those systems are designed to do just the opposite, for example the steps we take, including following the Euros, to downrank content that we think may be violating.

We know there is more to do and we'll keep making improvements. But regulation, focused on the systems and processes we have in place, has to be part of the solution. We've long called for new rules for the internet on areas like harmful content so that private companies aren't making so many important decisions on their own.

We're pleased the Online Safety Bill is moving forward, and in particular we welcome the Bill's attempt to establish a systems-focused framework, requiring service providers to use proportionate systems and processes to address harmful content, and the focus on increasing transparency through greater reporting obligations. It's vital that the final legislation helps make the internet a safer place, and is consistent and workable for the whole industry. So we will continue to work with the Government, Parliament and Ofcom as the Online Safety Bill passes into law over the next 12 months, to help make the new framework as effective as possible, and to bring additional oversight and confidence to the industry's work to keep users safe.

Yours sincerely,

Tara Hopkins
Director Public Policy EMEA, Instagram

Meta