



Joint Committee on Human Rights

Uncorrected oral evidence: [Democracy, free speech and freedom of association](#), HC 1890

Wednesday 1 May 2019

3.15 pm

[Watch the meeting](#)

Members present: Ms Harriet Harman (Chair); Fiona Bruce; Ms Karen Buck; Joanna Cherry; Baroness Hamwee; Baroness Lawrence of Clarendon; Jeremy Lefroy; Scott Mann; Lord Trimble.

Questions 19–31

Witnesses

[I](#): Ms Rebecca Stimson, UK Head of Public Policy, Facebook; Ms Katy Minshall, Head of Public Policy UK, Twitter.

Examination of Witnesses

Rebecca Stimson and Katy Minshall MP.

Q19 **Chair:** Welcome to this Committee session, and thank you very much to Rebecca Stimson, UK head of public policy at Facebook, and Katy Minshall, UK head of public policy at Twitter, for coming to give evidence to us today. We are the Joint Committee on Human Rights, which means we are half House of Commons and half House of Lords. As our name suggests, we are concerned about human rights.

In this inquiry—we have been taking evidence for some weeks now—we are looking at the balance between freedom of speech and expression on the one hand and on the other the right to life and safety of our democratically elected Members of Parliament and the importance of the protection of our democracy. Are we getting the balance right? Is there both freedom of expression and safety for Members of Parliament?

I think we will take as read the many benefits that social media provides for political discourse: the ability of MPs to communicate directly with



people in their constituencies, and more widely over certain issues, through Facebook and Twitter. I am afraid we are not focusing on the good side now; we are going straight to the problems, the damaging and dangerous side.

You will have seen all the published evidence of threats of violence and harassment of MPs. As part of our inquiry, this Committee has also been conducting confidential interviews with Members of Parliament, asking them about the nature and scale of the threats they receive and what they believe lies behind them. Many of them are concerned that these threats are being fuelled and incited on social media—on Facebook and Twitter—and that this is both generating the political and criminal threat and conveying it to the MPs. They are bombarded with personal threats to rape them, kill them, bomb them and their constituency offices and burn their houses; threats saying “I know where you live”, “We know where your children live”, “I hope your children die”; and messages with pictures of guns and nooses.

The context of all this is that we lost our Member of Parliament Jo Cox, who was murdered, and there was an attempted murder of another MP, Rosie Cooper. There have been many convictions of people committing criminal offences against MPs. I think we are all aware of the work done to deal with the threat of child pornography and child sex abuse via social media, and the discussion about how you tackle the incitement of terrorism on social media.

Are you aware that there is a strong view among MPs generally that what is happening on social media is a threat to democracy? They are prevented from engaging with their constituents in constituency surgeries in the way they normally do, they fear when they are in Westminster and away from their families that their children are vulnerable, they feel unable, because of the danger, to notify people in advance on Twitter that they are going to a particular place, and they change the way they do their advice surgeries.

In the next question, Joanna will ask you about some material on the screen. We know what you think about the threats of child sex abuse and how you are trying to deal with them, and the same with terrorism. How worried are you that our democracy is threatened by what goes on on social media?

Katy Minshall: First, I thank the Committee for inviting us to appear today. We are acutely aware of our responsibility and have come a long way. We are not where we want to be yet, and this is where the work of this Committee is incredibly helpful. Our strategy on the abuse of politicians has three elements.

The first is our partnership with the Parliamentary Security Department. As conveyed last week, the Parliamentary Security Department has access to a portal where it can expedite reports to Twitter. Its reports go to a specific team that reviews government reports. It can report a wider range of issues and give us more context. That process is supplemented



by a weekly phone meeting between me and the Parliamentary Security Department.

It is a really good opportunity to ask any questions and unblock any issues. We find it really helpful, because it can flag up things coming up in Parliament that may be controversial, and I can let the safety team know that we may see higher levels of reports. That process, which has been in place since the end of last year, has been a really useful step change. There are certainly aspects of it that I would look to replicate elsewhere. I am going up to Scotland in a few months and have already had preliminary discussions with some MSPs about what we might learn from it there.

The second part of this is changes to our product. It was unacceptable that we were wholly reliant on users to report issues to us. Now, 38% of our enforcement decisions on abuse are proactively flagged up by our own internal tools, and we want to continue moving the dial and reducing the burden on victims to report issues to us. Particularly relevant to politicians is identifying and removing what we call ban evaders: people we have previously removed from the platform who come back on to Twitter. We are getting better at detecting them. In the first three months of this year we removed over 100,000 ban evaders from the platform.

The final element of our strategy is what we can do offline. In partnership with PLAIT, the Parliamentary Liaison and Investigation Team—the Met team in Parliament—we have made ourselves available to all political parties for training on the safety features of Twitter. A number of organisations working on these issues are really good. One organisation, Glitch, runs training sessions for women who want to enter public life. I went to a workshop a couple of months ago. It gives information on the nature of online abuse and the safety features on social media to reduce the threat.

There is also something about being proactive with resources. Last week I was on the phone to DCMS, which is putting together a pack for candidates standing in the European elections to ensure that they have up-to-date information on Twitter safety.

That is where we are. I re-emphasise that I welcome the Committee's consideration of this and look forward to your recommendations on how we can do more.

Rebecca Stimson: First, thank you for the opportunity to speak here about these subjects. Similar to what Katie said, there are a number of things that we are doing that I am sure we will discuss in the course of this conversation. Facebook has had its community standards and rules in place ever since the beginning of Facebook, so for more than 15 years. Those cover our policies towards bullying, harassment, hate speech, public figures and so on. It is a very live document covering the nature of



threats, the things that we see and the feedback that we get. I, my team and others at Facebook are in and out of both Houses of Parliament almost every day, and we are well aware of the concerns that you outlined in your opening remarks.

We do a number of things. All the policies and tools that are available to the billions of people who use Facebook also cover public figures, so you can report anything to us 24 hours a day, seven days a week. We have recognised recently that there is a particular need to protect public figures, particularly in the UK at the moment, given the nature of the quite heated political debate, which might be a bit of an understatement. We have recently rolled out some changes to both our policies and our processes, so now we have a dedicated reporting channel for public figures in the UK to come direct to teams in our London office, where I am based, to flag things more quickly, as well as using the more than 30,000 people that we as a company now have working on safety and security. So that all applies, and all those channels are available to you.

We provide a lot of help and resources on how to use our tools, because what we do as a company is only one half of the coin; the other half is what you as individuals and your teams are able to do, and often we find that people need some extra help from us in working out how to block people, remove content and report content. So we have a range of training happening around that. Similar to Twitter, we work with the parliamentary authorities to deliver that alongside real-world training and support for how to maintain your own safety.

We have recently—maybe we will come on to this—recognised the changing nature of what is happening. We have made a number of changes. One, for example, is about recognising the particular vulnerability of female public figures and the targeting of messages and aggressive content at them. We have recently changed our policy to recognise that. Although swearing is allowed on our platform—we do not police for swearing—some terms are directed particularly frequently at women.

When that happens at scale on a daily basis, if you are a public figure, that creates an extremely intimidating environment that, as you said, might deter you from participating fully in the positive side of social media. Our change is that if you are a public figure, if those swear words are being used we will remove that content, even though in other scenarios that does not automatically violate our standards. Hopefully that shows you where we have been and some of the new steps that we have been taking to respond to the environment that we find ourselves in.

Chair: We will see some of those swear words, I am afraid. Sensitive viewers be warned.

Q20 **Joanna Cherry:** I was interested in what Rebecca said about Facebook



recognising the particular vulnerability of female public figures. I want to concentrate my questions on Twitter this afternoon. Does Twitter recognise the particular vulnerability of female public figures? We have heard evidence that female MPs come in for more abuse on Twitter than male MPs. Are you aware of that?

Katy Minshall: Yes.

Joanna Cherry: Are you aware of the Amnesty International research that was carried out in 2017 which found that Diane Abbott was the most abused female MP in the UK, and that contributing to that was not just misogyny but also racism? I should declare my interest in that I was one of the other top four of the most abused female MPs in the UK. So you recognise that there is a particular problem with the abuse of female MPs.

You will be aware that under the Equality Act sex is a protected characteristic, and that discrimination on the grounds of sex is unlawful. Why does Twitter's hateful conduct policy not include sex as a protected characteristic?

Katy Minshall: Our hateful conduct policy includes gender and gender identity as a protected characteristic.

Joanna Cherry: Yes, but under Section 11 of the Equality Act, sex is a protected characteristic. Why does your hateful conduct policy not protect women on the grounds of their sex? That is the question I am asking. Under British law—I hesitate to use that phrase because I am a Scots lawyer, but this is the law across the UK—it is unlawful to discriminate against people on the grounds of their sex.

I am puzzled as to why Twitter does not include sex as a protected characteristic. When we come to look at the very unpleasant videos and tweets that I am about to show, you will perhaps understand my concern that sex is not included as a protected characteristic by Twitter when in law it is a protected characteristic in the UK.

Can you tell us why Twitter has chosen to exclude from its hateful conduct policy sex as a protected characteristic when it is in law a protected characteristic? Could you give us an indication of the thinking behind that?

Katy Minshall: Our hateful conduct policy is based on UN definitions. You could not target someone who was female with abuse based on the fact that they are a woman; that would be against our rules.

Joanna Cherry: Are you saying that the inclusion of gender should be enough to protect women from discrimination by Twitter in the way it applies its policy?

Katy Minshall: As I said, there is a lot that we want to do in order to reduce the burden on reporters. We have rules in place that mean it would be a breach to target someone based on the fact that they are a



woman. We need to be far more proactive in reducing the burden on victims to report that.

Joanna Cherry: Clearly Twitter is operating across the globe. In the UK, sex is a protected characteristic. Are you able to tell us why Twitter has taken a specific decision to exclude sex as a protected characteristic? If not, could you perhaps follow up with a letter explaining the thinking behind that?

Katy Minshall: Yes, I am happy to follow up with a letter.

Joanna Cherry: I am afraid I am now going to show two rather unpleasant films attached to tweets. Before I do that, I want to make something crystal clear: these tweets use the word TERF, which stands for "trans-exclusionary radical feminist". They relate to the very active debate on social media about trans rights. I make clear that there is a considerable amount of abuse on both sides of the argument, and I deplore any abuse directed at trans people. However, in my questions today, because the evidence has shown that female MPs get the most abuse on Twitter, I want to concentrate on what Twitter is doing to protect women from misogyny on Twitter.

The first video that I am going to show was tweeted out by a Twitter account known as SonicFox and was initially ruled by Twitter not to be abusive or offensive, although I believe it was taken down eventually. I warn people that it is a video game or cartoon of a man repeatedly chopping a woman in the neck. I wonder if we could just look at it. The accompanying tweet reads "what I do to terfs". Are you aware of the controversy about this tweet?

Katy Minshall: Yes.

Joanna Cherry: You should have a couple of packs in front of you, pack A and pack B. On page 4 of pack A there is a tweet from a woman whose name I will not read out, because I am afraid that if I did she would come in for further abuse. This woman has tweeted a further tweet from SonicFox, who has retweeted Twitter saying that it received a complaint regarding his account for that film and investigated it but could not identify any violations of Twitter rules or applicable law and therefore the tweet would not be deleted.

Knowing there is a debate about trans rights, and knowing that some people on one side of the debate are called TERFs and that they tend to be women, would you not agree that that tweet was offensive on the grounds of sex, or perhaps even gender?

Katy Minshall: Yes, and that tweet was in violation of our policy and has since been removed. We have contacted the reporter to inform them that we made a mistake.

Joanna Cherry: When was it removed?



Katy Minshall: I believe it was last night.

Joanna Cherry: Was it removed because Caroline Criado-Perez, the well-known feminist, Helen Lewis, the associate editor of the *New Statesman*, and I had tweeted about it? Is that why it was removed—because an MP, a leading feminist and a leading journalist had tweeted about it?

Katy Minshall: I have been in touch with Helen and Caroline over email in the last couple of days. They made us aware of this issue.

Joanna Cherry: Were you aware that I had tweeted about it yesterday and would be raising it today?

Katy Minshall: I saw your tweet yesterday, yes.

Joanna Cherry: If a Member of Parliament, a leading feminist and respected commentator and a leading journalist had not tweeted about this, do you think SonicFox's tweet would have remained up?

Katy Minshall: That is an absolutely undesirable situation. I am due to meet Helen and Caroline in a couple of weeks, and I am aware that they have a number of issues that they would like to discuss with us. We absolutely want to get to a place where the situation that you outline does not happen.

Joanna Cherry: Let us look at the second film, which was sent to Helen Lewis when she raised concerns online about the first film. I warn people that it is a very unpleasant representation of a man flaying a woman alive. I could barely bear to watch it myself, so I am just warning those watching, both here and on television, that they may find it very distressing. Do you know what happened to the tweet containing that film?

Katy Minshall: I am not familiar with the specific tweet.

Joanna Cherry: Do you know if it was taken down?

Katy Minshall: I do not know if it was taken down.

Joanna Cherry: Do you think it should have been taken down?

Katy Minshall: Yes. I do not work in our safety team and am not an expert, but I think that that absolutely violates our hateful conduct policy.

Joanna Cherry: I am trying to understand why the first tweet, showing the chopping of the neck, was initially ruled all right by Twitter, and why it took the intervention of a leading journalist, a leading feminist commentator and a Member of Parliament for it to be ruled not all right.

In order to understand the answer to that question, we need to understand who is carrying out these decisions. Who is carrying out the



mediation at Twitter? Is it done in the UK or America? Where is it done, and who by? Is there any attempt at gender balance in the teams of people looking at these tweets?

Katy Minshall: Let me talk you through our training and how it works. There are three stages. The first is about how we develop these policies. We develop our rules in concert with a trust and safety council of organisations from across the world that have been in place for almost three years. They review our policy ideas and give us recommendations and feedback. That process has now been extended to public consultation, where appropriate. We ran a public consultation for the first time towards the end of last year on a new policy on dehumanisation.

When we have our policies, the second stage is when we onboard our content moderation team so that they can make decisions on content that they review. That includes training in the experience of members of protected categories, the cultural biases that they may see on the platform and the issues offline that vulnerable groups experience.

Joanna Cherry: Do you include training on discrimination against people on the grounds of sex? You see, you have not included sex as one of your protected characteristics, and I wonder if that could be what is going wrong here—the training is not covering the fact that sexist, misogynistic, demeaning behaviour should be treated as seriously as the abuse of, for example, trans people.

Katy Minshall: As gender and gender identity are included as a protected characteristic, my understanding is that it is absolutely included in the training. However, I am very happy to write to you afterwards with the details of that.

Joanna Cherry: Can I pursue with you whether there is any attempt at gender balance in the teams of people who carry out the mediation?

Katy Minshall: Again, I do not have the details, but I am happy to write to you afterwards.

Joanna Cherry: I wonder if we can look at another tweet that was sent to Helen Lewis of the *New Statesman*. It is on page 2 of pack A. Helen tweeted: "UPDATE: This tweet is officially not abusive. Thanks, Twitter. I will try to view it 'in the context of a larger conversation' whatever the hell that means". The tweet is a cartoon depiction of someone making the V sign—not V for victory but the other one—and holding up a gun and pointing it outwards. It says, and forgive my language, "SHUT THE FUCK UP TERF".

You will see that although it is a cartoon, the part of the cartoon that is a hand holding a gun is actually taken from a photograph, so it is a real hand holding a real gun. Do you agree that that adds to the sinister nature of this: the fact that it is a real hand holding a real gun?



Katy Minshall: Absolutely. The account that shared this photo has now been suspended.

Joanna Cherry: Yes, but initially, as Helen will tell you when she comes to meet you, Twitter said that this was not abusive and that she should try to view this “in the context of a larger conversation”. Do you accept that that decision was wrong?

Katy Minshall: Yes. We have contacted Helen, the reporter, and said that we made a mistake.

Joanna Cherry: There seem to be a number of mistakes here, and they seem to be mistakes that are failing to protect women. Would you accept that?

Katy Minshall: You mentioned the Amnesty report. I do not think anyone can read that report and not be horrified by the experiences that women in public life shared with Amnesty. Clearly there are a number of steps that we want and need to take, but we are in a different place from where we were even this time last year. There are a number of recommendations in the Amnesty report that we still need to move on, but we have already made three changes.

Joanna Cherry: My point is that there seems to be a pattern of Twitter initially ruling that extremely offensive, violent tweets directed at women in public life are acceptable, and that Twitter reviews its decision only when pressed by other figures in public life such as me, Helen and Caroline Criado-Perez. Do you not think there is something wrong there?

Katy Minshall: We are very much aware of the real issue that women experience on our platform. A few months ago, a colleague of mine from the trust and safety team met with women’s charities here in London, because we are reviewing our policy on harassment. There is an issue specific to women, typically ex-partners stalking them on Twitter in ways that have traditionally been difficult to detect under our rules, and we want to do better on that. We are acutely aware of the unique experience that women have on Twitter and the changes that we may have to make in our policies to get that right.

Joanna Cherry: Do you accept that the term “TERF” is a gendered term in the same way that—forgive me—“bitch” and “cunt” are gendered terms?

Katy Minshall: Yes.

Joanna Cherry: Are “bitch” and “cunt” acceptable on Twitter?

Katy Minshall: As Rebecca said, we do not have a prohibition on swearing, but there are a number of rules, so it would be a violation to target someone with that kind of language.



Joanna Cherry: So if someone sent me a tweet saying, "Shut the fuck up, cunt", would that be acceptable or would it be taken down?

Katy Minshall: As I said, I have colleagues who work in the safety team who would be able to advise on the specific nature of that tweet, and I am very happy to follow up with you in writing on that.

Joanna Cherry: Can you not answer that question? If someone sent an MP a tweet such as the one including the gun but saying "Shut the fuck up, cunt", would that be considered acceptable?

Katy Minshall: As I said, that tweet was taken down because it broke the rules, so in your scenario I would assume the same.

Joanna Cherry: Would you agree that Members of Parliament, women generally and indeed men ought to be able to discuss facts, evidence and debates in Parliament online without being shut down?

Katy Minshall: Yes.

Joanna Cherry: Well, let us look at pack B. Again, I am not going to read out the name of the person who tweeted in order to protect them from any problems that they might experience, but you will see that one person, using the hashtag #TwitterIsSexist, tweeted that Twitter suspended a woman from quoting MPs. The person who was suspended says, "When I was tweeting threads about the GRA"—that is, the Gender Recognition Act—"debates of 2003/4 I got locked out of my account mid thread. This happened twice". Why would Twitter lock someone out of their account for tweeting about what was said in Parliament during a debate on the Gender Recognition Bill in 2003-04?

Katy Minshall: I am afraid I am not familiar with this specific instance, so I am not able to comment on why.

Joanna Cherry: Would you be able to investigate for us and tell us why in a letter? You will see as I go through this that there is a pattern emerging here.

Can we move on to page 4? A woman here—I will not say her name to protect her identity, but this tweet is from November last year—said: "I was suspended from Twitter for saying we need to talk about male violence". Her tweet is addressed to a couple of other Twitter accounts, including Labour Left. It says: "We need to start talking about male violence. Males kill trans women. Males kill women. Males kill men. Males kill children. Most killers are male—vastly, hugely, disproportionately. Very, very few killers are female. Let's take action together to #StopMaleViolence".

You can take it from me, as someone who worked as a lawyer for years, that it is a fact that men kill more often than women—vastly more so. If



we assume that this woman was stating a fact about male violence, why would she be suspended from Twitter for that?

Katy Minshall: As I said, my colleagues in the safety team would have made this decision and I am happy to follow up in writing on the nature of that.

Joanna Cherry: On page 5, an account—again, I will not read out its name—says: “Twitter is giving out 7-day bans for quoting UK law on rape. This is outrageous censorship”. The person has said: “All rapists are men. In UK law, rape is a crime only committed by a person with a penis”. This has been said to be violating rules against hateful conduct. If that is a statement of fact, why would Twitter suspend someone for saying it?

Katy Minshall: A broader theme emerging from these examples is that working with safety organisations—committees such as this—in trying to think about how we can appropriately moderate a public discussion on an issue like transgender rights is difficult. We need to partner with people who are experts in the matter. With regard to that specific tweet, I would go to my colleagues, who would be able to provide the context for why that decision was made.

Joanna Cherry: There is no attack on trans people in this tweet. I emphasise this, because if there were an attack on trans people I would condemn the tweet. But the tweet says: “All rapists are men. In UK law, rape is a crime only committed by a person with a penis”. That does not seem to me to violate any rules against trans discrimination, which I would condemn.

On the last page, page 7, this account—again, I will not say its name—was determined to have violated the Twitter rules against hateful conduct by saying the following: “Over 80% of violent crime and 99% of sexual assaults are committed by males. You won’t find many women being kind when men try to take away what few safeguards they have. Sorry Philip, but we’ll be kind when men stop raping and killing us”.

Why does that violate the rules against hateful conduct? It is a bit of a generalisation—if I were a man, I would be a bit cross because I would think, “Not all men rape”—but Twitter is a bit of a limited medium for communication. There are factual statistics there, and I wonder why this seems to be immediately taken down when things such as “SHUT THE FUCK UP TERF” with a gun are taken down only after protests are made.

Katy Minshall: Absolutely. Clearly there is an issue here for us to look at. Cases like this are incredibly valuable in pointing out the limitations of our rules, and I am very happy to follow up with you in writing on these specific tweets.

Joanna Cherry: Can we be quite clear for the record that, despite the fact that Twitter has excluded sex from protected characteristics, it



believes that women should be protected from hateful and misogynistic conduct on its site?

Katy Minshall: Yes. As I have said, gender and gender identity are within our hateful conduct policy.

Q21 **Baroness Hamwee:** This question is for Rebecca. I am not a QC, so maybe you can relax to some degree.

You have already mentioned Facebook's community standards. I will read from them: "In determining whether a threat is credible, we may also consider additional information like a targeted person's public visibility and vulnerability", and I think you have mentioned MPs coming within the public visibility category. I appreciate that what we are all saying could be applied beyond the community of the cohort of MPs but this session is specifically about MPs and democracy.

The standards go on to say, "We remove content, disable accounts, and work with law enforcement when we believe there is a genuine risk of physical harm or direct threats to public safety". Why is this statement so limited to direct physical harm, genuine risk of physical harm and direct threats to public safety? A lot of other material breaches the law, mostly under the Public Order Act 1986. It is illegal intentionally to harass someone—we have heard that Twitter is reviewing its policy on harassment—cause alarm or distress, stir up hatred on the grounds of race, religion or sexual orientation, or to send malicious communications.

I am sure you know all this, but I am saying it so it is on the record. Why does none of that warrant taking down material when it is a part of our law?

Rebecca Stimson: There are a couple of things to say there. First, we have recently made a change to those standards. The part that you read out talked about credible threats of violence. I mentioned at the beginning the change that we have made on gendered cursing, recognising the particular vulnerability of female public figures. Even though swearing in itself does not violate our standards, we have changed how we react when a female public figure is exposed to that.

We have thought about—in fact we have not thought about it; we have actually done it—lowering the threshold for what we mean by threats. Our policy now regarding public figures, although I assume that our content standards have not been updated in the last couple of days to reflect this, is that it does not matter if you and I might look at something that someone has written and say that it was a turn of phrase or not a direct credible threat; anything that is a threat to a public figure, even if by any standards most people would not think that it was realistic, we will now remove.

Similar to the reasoning behind the change that we made regarding gendered cursing, we appreciate that when you are subject to that kind of language, even if it is not credible it creates an environment that is



intimidating. So our policy on threats to public figures is that we have lowered the threshold and will now remove much more of that.

Baroness Hamwee: So that is a lower bar than a person's "public visibility", which is the wording in the standards regarding a threat to a public figure.

Rebecca Stimson: The point I am making is that previously we would have judged whether to remove that threat by testing its credibility. We might have thought about whether the person had been specific in what they wrote. Did they say, "I'm going to turn up at this time and do this to this individual?" Obviously that would enhance what we might argue was the credibility of that threat. But if someone said, "If I could get my hands on that traffic warden, I'd wring his neck", I think most of us would agree that that was not really a credible threat.

Now, however, where that kind of language, that lower threshold, is directed at a public figure, we will remove it, even though most of us would agree that it was not really a genuine, credible threat. It is about the environment and the atmosphere around public figures, and that is what we are trying to help in bringing in that change.

Baroness Hamwee: A public figure's whereabouts are much more easily found than a traffic warden's.

Rebecca Stimson: Exactly.

Baroness Hamwee: I wrote down the word "credible", because you started by mentioning it. Should I be listening to those two points together?

Rebecca Stimson: Yes.

Baroness Hamwee: Right. Sorry, I interrupted you.

Rebecca Stimson: I think the other part of your question was about what gets taken down. Is that right?

Baroness Hamwee: It was about the other parts of the Public Order Act about stirring up hatred on the grounds that I read out.

Rebecca Stimson: Obviously anything that violates our standards—you have read some of them out, including our hate speech policies on bullying, harassment and so on—will come down. We use a mixture of human reviewers and machine learning.

As I mentioned at the beginning, we have 30,000 people now working on safety and security, and we have what I would argue is probably the most advanced automated system in the world to find some of this stuff. In our last transparency report—if you are not sure what that is, our community standards are available publicly—we report the statistics for on how well we are doing, which are increasingly good.



Baroness Hamwee: I was going to ask you about that, because the standards may be tweaked to take in the points you have just mentioned, but how does a member of the public using Facebook know that you also take seriously other breaches of the law?

Rebecca Stimson: You are absolutely right: our rules are only as good as our enforcement of them. That is why every six months we publish a publicly available transparency report, so that you can see how we are doing against some of our policies.

As Katy indicated, there are places where we are really good—on terrorism, child exploitation and that kind of thing. Our machines are able to find and remove around 99% of terrorist content before it is ever seen by anyone. It is much harder for a machine to identify accurately things such as bullying, harassment and some of the subjects we are discussing with you today. It might be us just having an argument about something using quite robust language.

We found about 2 million pieces of that kind of content, but only about 15% of that was found by our machines. For the rest, we rely on individuals reporting to us and human reviewers. It is often more about context and intent, which can involve more nuanced decisions. If you look at the statistics in that report—I am happy to send it to you afterwards—you can see that the trajectory, over the last few years that we have been reporting it, is positive all across the board. It is moving in the right direction.

Baroness Hamwee: So it relies on reaction rather than machine learning identifying everything that might be within the 2 million and then reviewing it. It is the other way around.

Rebecca Stimson: Yes, for some categories. It is hard to train a machine to work out what bullying is in every context.

Baroness Hamwee: I understand that. Just for clarity, you are saying that it is not a case of the machine identifying possibilities that come under bullying, for instance, and then a human looking at it. The human has to look first.

Rebecca Stimson: The machines are proactively finding 15% of what we take down for bullying. For the rest, at the moment you need a human being to look at it to make that judgment.

Q22 **Ms Karen Buck:** Flowing on very much from that is the question of the extent to which the machine learning and the moderation on both platforms are doing what they can—it will not be perfect—to understand specific contexts in different countries. We understand that some of the contexts specific to our culture and events would not necessarily be properly understood.

To what extent do the machine learning and the moderation reflect specific country factors, and how do you try to deal with that? You



already know the specific example of the Jo Cox murder. There will be posts—I have seen them many times—where people say, “Jo Cox had it coming”, “You’re going to get what Jo Cox had”, and things like that, which would not necessarily include words or images that in themselves would trigger moderation but would clearly have an intent. Do you do that? How country-specific is it?

Rebecca Stimson: It may be helpful if I spend a couple of seconds on how it works when you report something to us.

Ms Karen Buck: I was thinking not just about reporting but about you being proactive.

Rebecca Stimson: Sure. The way it works at the moment is that there is machine learning running on multiple categories of harmful content 24 hours a day. We roll that out across our platform only when it achieves a high level of accuracy in what it takes down, because we do not want to inadvertently censor lots of perfectly legitimate content, so there is a very high threshold before we allow that machine learning to operate. As I have said, it works extremely well for something readily identifiable such as terrorism, but not so well so far for bullying and harassment, but it will get there.

We have 15,000 human reviewers around the world looking at content 24 hours a day, seven days a week, in 50 different languages. Where content goes depends a bit on what it is. We have specialist teams to look at whether it is something to do with child abuse. It depends what language it is in and what time of day it is. Everyone might be asleep in the UK, and if you report something or something gets flagged up to us, it needs to be seen by someone.

Part of the training for our human reviewers is for contexts in which they think, “I’m a human reviewer in another country and I don’t even know who Jo Cox is”. They will flag that up to be dealt with by our content moderation team in Dublin, which will have that more local knowledge and context.

Ms Karen Buck: If I am interpreting you correctly, that would rely on somebody reporting it. If tonight I get a message saying, “You’re going to get what Jo Cox got”, I have to report it; there is no other means of any of your systems necessarily detecting that.

Rebecca Stimson: If it were attacking you on a protected characteristic or included clear threats of violence—

Ms Karen Buck: But in the specific example I am using—

Rebecca Stimson: If it just said, “You’re going to get what Jo Cox got”, I do not believe our systems are sophisticated enough yet to work it out. You and I would understand what that meant; the machines cannot at the moment. If you then flagged it up or if someone saw it on your page



and flagged it up to us, our reviewers would look at that and it would come down.

Ms Karen Buck: What about Twitter?

Katy Minshall: It is similar for us; the technology is not there yet. This is where partnerships are so important, in two ways. We had an example a couple of months ago where the Parliamentary Security Department had reported something to Twitter and it had come back as not a rule violation. It was a niche term that our content moderator was not aware of as an issue. He reported that to us and we made the change and shared that with our team. It is a live and constantly evolving effort.

The other side of this is proactive work. The Parliamentary Security Department proactively monitors the accounts of Members of Parliament who have signed up. So while the technology is not there yet, there are things we can do to reduce the burden on politicians having to identify and report this content.

Ms Karen Buck: How do you respond to a second category, the pile-on? We have talked about specific threats and specific abuse, but it is equally—sometime more—horrific to be on the receiving end of 1,000 or 1,500 abusive messages, even if none of those individual messages threatens violence. Is there a way of knowing what is going on, particularly for public figures? If so, what can be done about it?

Katy Minshall: Mob harassment is against our rules. That is as applicable online as offline. As in the offline world, you cannot go on Twitter and say, "Let's all go and harass this individual". We use technology to detect mob harassment. I was struck when the Law Commission review recommended something in the criminal law targeting that specific issue more effectively. I welcome its consideration of that.

Ms Karen Buck: What about Facebook?

Rebecca Stimson: Similarly, it would fall foul of our harassment policies. I do not believe that we proactively detect for it at the moment. It is worth remembering that Facebook is a slightly different platform and works in a different way; it is predominantly about friends and family, and there is more narrow engagement on MPs' pages.

I do not believe we currently have a way of automatically detecting if that is happening. If it were flagged to us, it would fall foul of our harassment policies.

Ms Karen Buck: This would effectively be if X person—by definition, probably someone with a high profile and number of followers—were to say to their followers, "Let's all weigh in on this person". Give me a form of words that might attract that kind of reaction from you.



Katy Minshall: I do not think it would be a form of words but behaviours: lots of people suddenly tweeting at a user, particularly if that user did not follow them back. Particularly in the context of politicians, lots of bystanders make reports to us—we review hundreds of reports about MPs—so the combination of the technology looking for that and the reports means that that kind of review would be prioritised and mob harassment detected.

Rebecca Stimson: One of the other relevant things here is that when you see this kind of thing at scale, it is often co-ordinated fake accounts and bots. We have a strong policy and an ability to detect and remove fake accounts. You can see odd behaviour, such as an account contacting someone 1,000 times in a minute; you can spot signals. That can often be associated with the kind of problem you are describing.

Q23 **Scott Mann:** On that point, I want to flush out how you review closed groups. Some of the biggest problems with the dark web are these groups that just formulate. I know Facebook has the option to close groups off. Do you have a human reviewer who can go in and check those groups? Is there a higher level of scrutiny for some of those closed groups?

Rebecca Stimson: The first point to remember is that they are not closed to us. We can see absolutely everything that happens in those groups; the AI tools are constantly scanning them all. By making yourself a private group, you do not remove yourself from our rules or from the tools. People were quite concerned—it was disconcerting, especially for public figures, that you cannot see it unless you are a member of that group—so we have recently made some changes to try to go even further. Now, for example, if a group regularly posts content that violates our standards, we will hold the admins of the group accountable, so you cannot just preside over a group that has this kind of content in it.

We will flag up groups that have had violations and are probably a bit low-quality and not great, to deter people from following them. We will also downrank them, which means that if you search for a particular thing or happen to follow a certain set of interests, we are making groups that are considered poor-quality and have violations against them much harder to find. Downranking is actually very effective; there are lots of studies in the US that look at some of the statistics on downranking poor-quality news and show that it gets 80% less engagement, so it really drops off. Those are some of the things we are doing.

Some people ask, "Why do you bother having these things? Aren't they terrible?" The vast majority of them are used for really valuable and legitimate cases, such as people discussing medical conditions or campaigning for LGBT rights in countries where that might be challenging. It is always worth remembering that they have a value, but the key point is that they are not closed to us and that everything I have described applies to them.



Q24 Jeremy Lefroy: I will now move on to the material reported to the police. Do you have any figures on what proportion of material you report to the police?

Katy Minshall: Let me outline how we work with the police in general. There are a few points in the context of this discussion. The first is that we have a specific online portal for any law enforcement body to report issues to us or to ask for information.

That is coupled with an online submission site that we launched a few years ago for whitelisted police forces or other law enforcement government agencies that are regularly downloading data and making information requests, removal requests or preservation orders. We have 68 different email domains in the UK onboarded to that system, which will represent thousands and thousands of people.

That process is supplemented with training. We regularly run training for the police in how to make requests for data to us. We had one in London in March, including a meeting with the College of Policing on how we can scale that so that all officers have the understanding they need to make requests to Twitter.

The second point is that I am a point of contact for PLAIT and the Parliamentary Security Department, and we sit down regularly. Sometimes it is helpful for them to pick up the phone or drop me an email with a quick question. I find it useful in case there is an emergency in Parliament.

There is also something for users. We give users who report a violent threat to Twitter the option to generate an email. In that email will be a copy of the tweet and instructions for law enforcement such that you can go to an officer with your smartphone or a printout and everything is there for the police officer to take that forward.

Rebecca Stimson: Similarly, we have a law enforcement team in the Facebook London office, heavily staffed by people who were previously in UK law enforcement. We have a whole span of relationships across the intelligence communities, the Home Office and law enforcement directly. Obviously, under the legal framework, law enforcement has the ability to request information from us. Our next transparency report, which comes out soon, will show that last year we had just over 15,000 requests and complied with over 90% of those.

That covers what they have requested from us. As Katy has said, in severe cases, where we see immediate emergency situations or threats of real-world harm, we will also proactively report back to the police and encourage our users to do so.

Jeremy Lefroy: This next question comes out of conversations with my local police force in Staffordshire and is not specific to Twitter or Facebook. The amount of work that has been generated by social media



for police has put enormous strains on the police force that were not really there before. Do you both recognise that this is additional work? It is not due to either of you two; it is due to people who use social networks, not just your two. Is that something that you have heard—that this has imposed a pretty large burden on the law enforcement agencies generally?

Rebecca Stimson: We have absolutely heard that. You may be aware that Facebook was in front of the Home Affairs Select Committee last week—as was Katy, who is doing the rounds this week—and one of the conversations that they had in that context was about reporting to the police. My team was then in the serious crime meeting chaired by the Home Secretary immediately after that and got the message very clearly: “Do not send us every single possible thing that you might see on your platform because we simply couldn’t deal with that volume”.

As I have described, the way that it works—reasonably well, although it could be better—is that we set quite a high threshold, at “This is an imminent, serious, credible situation that is live and we must flag that to the police”. They, with their vastly superior powers of investigation, can then direct us through the legal framework to request information for what else they want and what they find useful. We talk with them and have an ongoing relationship, as I said, on a daily basis through our law enforcement team about what else is helpful and how we do that.

In the US they passed a law that anything that we and other platforms see on child exploitation had to be passed to the police. Facebook designed and developed a system for prioritising that content, because there was so much of it that we had to help law enforcement to filter through harmless innocent material that our systems had flagged but that turned out to be nothing and serious material that needed looking at immediately.

Where we have responsibilities to hand that information over, we have demonstrated that we are happy to help them to deal with the potential volume of what arises. At the moment in the UK, that is done more by conversation between us about how we share information.

Jeremy Lefroy: That is very interesting. So you and the police between you have a way of assessing the seriousness.

Rebecca Stimson: Yes. This is quite new territory. A lot of challenges have been thrown up by social media and the internet generally, and lots of platforms and internet companies are in these large discussions, hosted mostly by the Home Office, about navigating through exactly the challenge that you are raising.

Jeremy Lefroy: You see on some press websites comments that have pre-moderation. With the kind of content that we are talking about, that would clearly be difficult, but do you have any means whereby, for instance, there could be a time delay, of seconds or perhaps a minute or



so, before someone posted, which would allow your systems—your algorithms and so on—to check whether something was not in compliance with your rules and therefore could be prevented from being posted even before it had been, rather than being taken down in respect? This question is for both of you, but I am talking to Katy first here.

Katy Minshall: The comparison with the press is really interesting. The White Paper contains some *Guardian* research that was undertaken, which looked at 17 million comments and saw that of the top 10 most abused writers, eight were female and two were black men. So this is not an issue that is unique to social media; it is absolutely reflected across the internet. To come back to your question on—

Jeremy Lefroy: Time delay.

Katy Minshall: On Twitter we see 500 million tweets a day. Pre-moderation would be extremely difficult to achieve.

Jeremy Lefroy: I am not talking about it being done personally. I am talking about it being done through AI, in effect. Clearly it is impossible for to be done by individuals.

Katy Minshall: Our proactive tools looking for abuse kind of work in that way. They are scanning when tweets are uploaded to Twitter. That is the direction that we are heading in.

Jeremy Lefroy: But at the moment—and forgive me if I have got this wrong; I know very little about this area—the tweet would appear and then maybe get taken down. I am asking if you are working towards a situation where in some cases it would not even appear, because your tools will have prevented it from appearing because the characteristics are set at a fairly high bar.

Katy Minshall: Yes. We are probably in a similar position. With things such as terrorist content, we are in that space. If something is uploaded that we have previously identified as being an image that is within our database of content that is illegal, that tweet—

Rebecca Stimson: Some of the figures that I have been given on what we find—3 million pieces of terror content, 99% proactively—show that it is doing exactly what you are describing. Our systems are so quick that as someone presses “Post”, they catch it and it never appears at all.

As I said, we are very good at that in some contexts and still working on it in others. If you are looking at doing that at scale, as I said in a previous answer you have to think about the accuracy of what you are doing. People’s expectations are that they are not being pre-screened and pre-read and that they are allowed to post within the rules on Facebook as they wish. There would be some questions about speech and privacy if you were suggesting large-scale pre-screening, which I know you are



not; you are talking about that kind of machine learning in order to capture something instantaneously, and that is happening at scale now.

As I have said, we have different kinds of technology. For example, there is a thing called hashing. Where there is an offending photo or video, hashing is like putting a fingerprint on it. Once you have done that and you bank it, no one can upload it; the machines will catch it immediately if anyone then tries to re-upload that image.

So some of that technology is already doing exactly what you are describing but, as I have said, in the context of something more nuanced and difficult such as bullying, it is much harder. I suspect that there will always be some need for human review, and that will inherently slow down the process.

Katy Minshall: The Law Commission review talked about the ambiguity of terms such as “grossly offensive”. A machine would really struggle to make those determinations when the courts here already struggle to make them.

Jeremy Lefroy: I understand that. What about the legal situation? Clearly the law is struggling to catch up with social media. This has probably been asked by other committees, but are there any recommendations that you would make about both the criminal offences being committed and civil offences such as libel or slander?

Katy Minshall: I think it is incredibly difficult. Only 3% of recorded malicious communications result in a charge, and that is online and offline. We will move the dial if there are real-world consequences to this sort of behaviour.

What is really challenging is the volume of what we are taking down, the vast majority of which would not be illegal. Commander Usher told this Committee last week—I will follow up with the exact wording, because I do not want to misquote him—something to the effect that in their investigations it has not necessarily been the person who is tweeting online abuse that they worry about as much as the person who may be flying under the radar and who, if they are on Twitter, may “like” that tweet.

What we do next with regard to the utility of us providing data is very much a question for the police. It is a complex legal question. As I am sure has come across in this discussion, we work very closely with the police but we also have legal obligations under privacy law. There are all sorts of contradictions that we need to work through.

Q25 **Jeremy Lefroy:** Lastly, there is a whole question of—I cannot think of any other phrase for it—good behaviour or politeness in discourse, which has perhaps been ignored too much; it has just been assumed that it will happen.

What work are both your companies doing to ensure moderate, polite and



reasonable discourse that does not use swear words? On the one hand we want to encourage vigorous debate, but on the other hand we want—at least most of us do, not just in this room but elsewhere—to ensure that it is done in a reasonable manner, vigorous but polite, that uses language properly, not offensively.

Rebecca Stimson: I have already talked about the change with regard to gendered swearing in some of my previous answers and recognising that, while swearing is permitted on our platform, in certain circumstances there are particular words that are repeatedly directed at female public figures, so even if the rest of what someone is saying might just, as you say, be a very robust comment about a particular situation, that material will now come down because of the environment and the atmosphere that it creates.

Your question is very difficult for us. I think it would be the wrong policy decision for us to ban swearing; that does not feel proportionate. This is why we give all of you who have pages on Facebook tools such as the profanity filter. People are quite subjective about this issue; some do not mind it in the slightest while some find it very difficult. You can put in the words that you do not want to see, those filters will then apply to your page and that language will not appear.

Katy Minshall: Our CEO has spoken about this a number of times. We are looking at everything, including the fundamentals of our platform. We have a partnership with the University of Oxford, which is looking at how exposure on Twitter to a variety of different viewpoints can decrease prejudice.

There are a couple of other points that I would make. The first is how we use signals to consider how to organise and present tweets in communal areas, such as search and comments. We now take into account a number of signals, such as whether a user is tweeting, at a very high frequency, lots of accounts that do not follow them back. That is probably a signal that they are not engaging in public discussion in good faith. So that user may be hidden; when you see comments below your own tweet, his reply may be in the “View more replies” section. Using that technology has already yielded results; we have seen a drop in abuse of 4% on search and 8% on conversation. The more that we can think about that, the better.

There is also the question of user education. We at Twitter have a system that we call the penalty box. Where appropriate, if a user violates our rules they will be placed in a time out. They will be told what rule they broke and the tweet that broke it, and they will have to click through in order to reactivate their Twitter account. We find that the majority of Twitter users who are placed in that state are there just once. They learn from their behaviour and they change it, and they do not break the rules again.

Q26 **Chair:** We have talked about a new zone, with the advent of social



media, for the conduct of criminal activity: threats to kill, which are criminal; incitement to racial hatred, which is criminal; and other matters that Sally referred to.

To follow up on Jeremy's point about policing and resources, if there is a football match and the police are needed in order to deal with public safety or avert threats of violence, the football club pays for that policing because it is generating the potential threat, so under the law it is required to pay when it uses the police's services.

Bearing in mind that you are profit-making organisations, have you thought about why the public sector should pay for the criminality that is finding a home in your companies? Why should the public purse pay for that? I will leave aside how much tax you pay and what the proportionality is there, because that is another argument. Has that been raised with you?

Rebecca Stimson: We have described the way we currently work with law enforcement, what we report to them and the legal framework for them to request information from us. It is always worth remembering that we are talking about what one user says to another, and that sometimes there are quite complicated decisions to be made about whether something constitutes criminality or not.

That is why we have this ongoing relationship with law enforcement. I have referenced the example in the US where we were mandated by law enforcement to do something at a very significant scale. We were very happy in that instance to step up and provide them with software that we designed, developed and rolled out to them to help them to prioritise what they were seeing.

It is an ongoing conversation, and we absolutely recognise that, particularly given the scale of Facebook. We have 40 million active daily users in the UK. Few of them are engaging in anything that anyone would regard as criminal activity, thankfully, but that is why we have this conversation with police and law enforcement: to work with them in the legal framework that exists and see what we can do to help and support them, up to and including what we did in the US, providing them with resource from our expertise to help them to process that information.

Katy Minshall: We will always look for ways to work effectively with law enforcement. Government reports constitute very few of our removals. We review over 200,000 reports a day, and government requests constitute a very small proportion of those.

Chair: But do you not see the analogy with football clubs? If you are hosting criminality, you should pay for the policing of it. You are making a profit out of it—the enterprise, not the criminality—but a by-product is that criminality is happening on your site. Should you calculate how much it costs the police to deal with that?



Rebecca Stimson: This is potentially a conversation to have in the context of the online harms White Paper. The Government have set out the harmful behaviour that they are worried about. This is all about user-generated content—what one person is saying or doing to another—and we are fully engaged in the conversation.

I suggest that we do not see large-scale criminal activity on our platform. We take down hundreds of millions of reports every day, and we had 15,000 from law enforcement in the whole of last year, so it is worth thinking through the scale. However, if the Government find that they wish to describe some form of criminality that is identifiable on our platform, and if it is an issue in the context of the White Paper discussion, we will be happy to work with them on that as we consider the harms and possible solutions.

Q27 **Scott Mann:** My first question is to Katy. I want to understand your appropriateness rules and how someone makes that call. I understand that the algorithm might like something or pull it out. When it does, is there one individual moderator who makes the decision about whether or not that is an appropriate tweet, or is there a team of people who discuss that particular tweet and work out if they feel that it is appropriate?

To give you an example from some of the evidence that we have had, one of our MPs had a death threat that was considered to be an opinion rather than a full-blown threat; there were tweets about Novichok that were not considered to be threatening. I want to drill down on how that decision is made.

Katy Minshall: When you say “appropriateness”, do you mean whether or not they break our rules?

Scott Mann: Yes.

Katy Minshall: A member of our content moderation team will review a tweet. If they are unsure or if they have questions, that may be escalated for further review. It may be passed on to another team. For example, we have a dedicated team for crime, so if the question concerns something illegal, they may pass it on there. There is no hard-and-fast rule for what happens with every tweet.

Q28 **Scott Mann:** Okay. I have a second question to both of you. We were told last week that the Parliamentary Security Department has the ability to be a trusted partner and can remove content quickly, but MPs do not currently have the ability to do so themselves. Is that something that you think we should have the ability to do?

Katy Minshall: There is no reason why we cannot explore something similar for MPs, but the reason for the PSD playing that role was to reduce the burden on politicians themselves having to look for that content. The PSD proactively monitors social media accounts and reaches out to us to remove content.



Rebecca Stimson: On your personal pages and your pages as MPs, you already have the ability to delete, remove and block people as you wish. As Katy said, we added that extra layer as a trusted partner with the authorities so that they could help you with that task and flag stuff so that we could remove it directly.

Scott Mann: One of the biggest issues that I have with both your companies is that you allow people to be anonymous online. Why would you do that? What is the purpose of having anonymity on a site that is socially outward-facing? We see that when people hide behind a screen they tend to be a bit braver than they are in real life. From my point of view, I cannot understand why companies allow people to have that anonymity.

Katy Minshall: There are a few issues here. Before I get into them, though, I want to say very clearly that yes, you can be pseudonymous but you cannot have a fake account. I could not have a Twitter account with a photo of Rebecca saying "I am Rebecca Stimson". That would break our rules. We can see a large amount of information about a user that we can work with law enforcement to provide, beyond whether they have said their name or not.

To answer your question, there are three points about verifying identity: do you want to do it, how does it work, and will it work? First, do you want to do it? There are real trade-offs in requiring people to declare their identity.

Chair: To follow up Scott's question, are you talking about declaring it publicly?

Katy Minshall: Declaring it on Twitter. Something that was trending on the platform towards the end of last year was the hashtag #WhyIDidntReport, which was individuals sharing stories of why they had not reported sexual harassment, many of which were anonymous. There have been a number of occasions on Twitter where individuals in war zones have shared imagery of what is going on on the ground, and having to declare their identity would be a barrier to that. So there are trade-offs in doing that.

However, let us say that we are going to require identity none the less. Who does that? Do people feel comfortable with Twitter now asking for personal information about individuals? Historically we have had a data-minimisation approach. How do you do it? What ID do you ask for? We do not have ID cards in the UK, so are we going to use passports for verification? That could exclude some of the more vulnerable people in the UK who could not verify their identity and be on a platform like Twitter.

Then there is the big question: if the aim is to reduce online abuse, will it work? There are not many case studies that we can look at. There is the example of South Korea, where the results were mixed at best. I would



say, from my experience of looking at Twitter abuse every day as part of my job, that there are plenty of occasions where people do not abuse with anonymity. They display their name, more than their name—they have in their bio information about where they live, their job and all sorts of indicators.

If we are going to go down the path of removing anonymity, we should be absolutely sure that it will work, and I am not quite clear whether that is there yet.

Rebecca Stimson: Facebook has a real-name policy, although it is not foolproof. When you sign up to Facebook, you have to use an email and, I believe, a phone number. As with what Katy was saying, that is not watertight, but we do not allow people to have an obviously fake name such as a nickname. It is a predominantly real-name policy.

Facebook is a very different kind of platform from Twitter; it is much more about your friends and family and less about public content. To echo one of Katy's points, we actually get asked why you cannot have anonymous profiles on Facebook, such as for people who live in a community where it is difficult to be LGBT; they want to be on the platform to talk to their friends but do not want their mum or dad to see.

There are a few cases where we get asked the opposite question: why we do not allow anonymous or nickname accounts on the platform but instead they have to be under a real name.

Scott Mann: Have you ever considered a code of conduct for people to sign up to when they enter your platform, in order to avoid abuse? It would mean that they knew when they were signing up that at any time their account could be completely withdrawn if they broke the terms that you set for your platforms.

Katy Minshall: That is what we do at Twitter. We have terms of service that you sign up to when you create a Twitter account which make very clear that we reserve the right not only to remove you from the platform if you break the rules but to change those rules.

Rebecca Stimson: As I said, our content standards have been public for a very long time. People understand what those rules are. As we all know, informed consent on the part of users is still a bit of a challenge. We have found that people might not pay a lot of attention to those rules at the beginning, but as soon as something goes wrong they will read them in great depth.

Scott Mann: It focuses minds.

Rebecca Stimson: It does.

Q29 **Lord Trimble:** I move on to something slightly different: political debate and discussion. That is legitimate; it is necessary that people should be



able to criticise political measures and people involved in politics. The question is how you strike a balance and identify situations where that is desirable and something that in many respects is to be encouraged. Do you have a higher threshold than you would in other matters?

Katy Minshall: We very much take the view that freedom of expression is meaningless if people are silenced and afraid to speak up. Safety and freedom of expression are often presented as mutually exclusive, and there are indeed tensions between them that we have to work through. For us, though, safety is inherent to freedom of expression because we want everyone to feel safe in expressing their unique point of view.

Rebecca Stimson: Similarly, Facebook is not a free-speech platform. We have had our rules for 15 years now. Your point is right: we recognised at the outset of this conversation the incredibly valuable role that social media can play in facilitating these conversations.

Where exactly to draw the line between bullying and harassment and hate speech is sometimes very clear but sometimes really difficult. That is why we talk every single day about our standards and pieces of content, just as we have done today. That line is very hard to strike, which is why we work with lots of organisations outside ours to define our rules and constantly refine them. I have mentioned the example of our public-figure policies that we have updated recently to reflect the feedback that we have heard from you and your colleagues.

Chair: Do you have a higher threshold in these discussions with people who have stood for election to public office and put themselves in the public domain? Do you expect them to take more abuse or threats because they have put themselves in the public domain? Is there a higher threshold for them?

Rebecca Stimson: No.

Q30 **Baroness Lawrence of Clarendon:** My question is about MPs who have said they do not want to continue being an MP. When they call out abuse that is happening, they feel that that triggers more abuse. Are you confident that your platforms are not undermining democracy when that happens?

Katy Minshall: We are doing our own analysis of this issue. That is ongoing, but our preliminary finding is that there is a real issue with content that does not break the rules but is relentlessly negative and could push someone to think, "I don't want to stand for election and be in that space any more".

There are things that we can do. In June we are going to start trialling auto-moderated replies in one of the countries that Twitter is in, so that if you tweet something and you get a load of comments below it, you will have the ability to hide comments that you do not want to be there. That is only a trial. There are all sorts of positive applications of it but also



opportunities for people to use it with bad intentions, so we will see what happens. I will keep the Committee updated.

Rebecca Stimson: As we recognised at the beginning, almost every MP has a page. Lots of surveys have shown how incredibly game-changing that has been, in their conversations and their ability to communicate. Nearly 2.5 million people specifically follow an MP. So it is easy enough to say, "I think it's a force for good generally", but today we are focusing on all the issues that are highly damaging. It would be awful if anyone felt too intimidated to run for public office because of this kind of abuse.

We have all our policies and, as I said, the investment that we are making in moderators and AI to try to address that. We have recently taken down a swathe of accounts of far-right groups in this country. They were consistently violating our policies and tipped over into organised hatred. It was not just a political ideology that might go up to the line of what you think is acceptable; they clearly tipped over it.

In those contexts, we also take into account the real-world information that we will get from law enforcement and other communities about things that are happening away from our platform to inform that decision. So we are trying to create that better atmosphere and remove some of the more extreme behaviour.

Katy's reference is right: a lot of this is about the users and the language that people currently feel might be acceptable. Both companies invest a lot of time and effort in programmes with young people and adults about digital literacy to try to improve that kind of discourse and think about the impact that even things that do not violate our policies might be having on public figures.

Baroness Lawrence of Clarendon: In the last couple of years, that has got even worse, especially in today's climate where MPs feel really threatened. Twitter says: "Free expression is a human right—we believe that everyone has a voice, and the right to use it".

Many MPs no longer use social media because of the scale of the abuse. What is the effect on social media's usefulness as a forum for democratic debate if policy formers are not able to take part in the debate?

Katy Minshall: It would be highly undesirable if MPs stopped using Twitter. MPs have a combined base of 17 million followers, and in the eight months that I have been at Twitter I have seen some incredible and innovative uses of social media to connect with constituents and to create awareness and shine a light on new issues, so it would be immensely undesirable for MPs to stop using the platform.

Baroness Lawrence of Clarendon: Do you have anything to add on that, Rebecca?

Rebecca Stimson: Katy echoes what I have said. We see great value in the use of our platform in democracy. Free speech and the exercise of



human rights are the subject of this inquiry, which is why it is such a priority for us not only to tackle the kind of abuse that you are seeing but to ensure that you as users, but also as private citizens, have the best and most positive experience online.

Even if you disregard our sense of responsibility towards those subjects, it is bad for business as well. No one wants to see people thinking that it is not a nice experience for anyone to be there. We want our users to stay and use our services. That is why we are investing so heavily in trying to address these problems.

- Q31 **Fiona Bruce:** I would like to ask you about candidates, people standing for public office as MPs, MEPs and indeed councillors; thousands of candidates will be standing for election tomorrow. Do you agree that, during the period when they are candidates, when they are campaigning and their leaflets might be going through thousands of doors in their locality, they are public figures?

Katy Minshall: Yes. This is a really key issue. We have jointly made contact with the parties to try to be at candidate training weekends to make sure that we are giving them not only a point of contact but the fundamentals of safety features on our platforms. Many candidates create a Twitter account when they become a candidate, and it is critical that we are able to reach them.

Rebecca Stimson: It is the same for Facebook. Particularly around a large or key vote, we will ramp up our efforts across the board to reflect the particular issue that is happening.

Fiona Bruce: Do you think there are going to be particular problems in the next few weeks as we head into potentially highly contentious European elections? During these periods of campaigning, figures come under intense scrutiny and things can become very heated. I give you the offline example of hustings, when candidates are often in public halls. This happened to me three or four years ago: the debate got heated, and a man jumped on the platform and hit me so strongly that I was knocked over and he had to be arrested by the police. When I talked to the police a few days later, they felt that the atmosphere in the room had stirred him up and provoked him.

Do you anticipate that in this European election there could well be a problem with social media generating hatred and violence during the campaign? What are you doing about that? You talked about contacting the parties, but obviously in going to events you will only ever have contact with a handful of the potential candidates who are going to stand. For council elections we are talking about thousands of candidates.

Katy Minshall: What I think would be a really good step forward—certainly during elections, but there is no reason why it could not be a standing measure—is regular meetings and joint work between social media, politicians, police and traditional media organisations. There was a



documentary on ITV a couple of months ago in which a researcher shared that official online hate figures had spiked by 20% after the infamous “mutineers” headline. Martin Hewitt, the chair of the NPCC, put it really well when he said that it behoves all of us to do what we can to ensure, particularly during elections, that the nature of the debate is as good as it can be.

Fiona Bruce: Would you like to add anything, Rebecca? Obviously many people will view this hearing who may well be candidates. Is there anything that you would like to put on record by way of advice to people who are standing in elections right now?

Rebecca Stimson: On your first point, all I would add is that, as you say, we are ramping up our efforts for people running in the elections tomorrow and anyone who may run if the European elections go ahead.

All the tools and resources that I have talked about throughout this session are available to those people during those periods of time. We have two election operation centres, one in Dublin and one in Singapore, that run the entire time, working on major elections around the world. When there is a major vote in a country—let us say, for example, that the European elections go ahead—there is a team that comes in to bolster all the things that I have already described, with 20 or 30 people coming in to try to address the range of issues that could crop up, such as dealing with content problems very quickly.

There is a multitude of things that we do. We are able to tackle disinformation and fake news that might be being spread on our platform.

We have rolled out a whole new load of changes with regard to political advertising as well so that people can be clearer. We draw that definition quite broadly; it is not just about the main parties, or just in an election period. If people are running political ads that refer to a political figure or issue, those are now labelled with who has paid for them, and there is an archive behind them so that you can see what other adverts they are running.

So there are multiple measures that we are taking that play into your question. We are taking them because, as Katy said, we recognise that the upcoming European elections in particular, because of the nature of the conversation, could be very challenging and fraught for the people taking part.

Fiona Bruce: You mentioned a DCMS pack. How are you inputting into that?

Katy Minshall: As I understand it, DCMS is putting together a pack for candidates standing in the European elections. We spoke last week to ensure that they had official Twitter safety advice that they could include



in the pack. I am happy to put you in touch with the official who is working on that.

Fiona Bruce: I am sure we could do that. If a candidate standing now is concerned, how do they contact each of your organisations very quickly and easily for the advice that you have?

Rebecca Stimson: We have a website that becomes available. As you say, it is a big challenge in the local elections; something like 20,000 councillors are running, so the service has to be slightly different. We have a point-of-contact website that they can contact, and then we will pick up whatever they are concerned about and trying to flag to us.

Katy Minshall: We have shared our contact information as part of the DCMS pack. We also have a bespoke inbox for government questions.

Baroness Hamwee: Are you in touch with the Electoral Commission about these issues?

Rebecca Stimson: Yes.

Chair: Are you anticipating that, in the European elections, the general climate and heightened atmosphere that Fiona referred to will precipitate more threats against Members of Parliament finding their home on Facebook and Twitter, as well as against candidates in the European parliamentary elections, because of the debate around that time? Is that what you have already thought through and worked out? If so, what do you plan to do about it?

Rebecca Stimson: As I said, when there are major votes in any country around the world, including here, we definitely see much more debate and dialogue going on. There is much more interaction with candidates; more information is put out and people will respond to it. So we are already well ahead in contingency planning, including talking to the Electoral Commission, about what to expect and what may happen.

When we get near to that election, we provide a huge amount of resources and support to candidates, and then we ramp up our teams based here in London to ensure that we are available as quickly as possible to deal with issues as and when they come up.

Chair: I am referring not just to the threats to candidates, which Fiona raised, but to the threats to MPs, which will also ramp up, although we are not candidates in the European parliamentary elections. What about threats of violence and illegality against MPs?

Rebecca Stimson: Sorry, maybe I was not clear. We are treating that as the same thing. I mentioned in one of my earlier answers that we have already rolled out a designated channel for existing MPs to raise particular concerns with us. That will be available throughout if the election goes ahead next month.



Chair: But are you ramping up your surveillance with regard to what is happening to MPs at a time when they will be particularly under threat because of the heightened atmosphere of the European elections?

Rebecca Stimson: We have a dedicated politics and government team in London who deal specifically with these issues. We have put a dedicated reporting channel in place, which I just referred to, and we are working closely with the parliamentary authorities. So yes, we are ramping up our efforts in relation to what may happen.

Chair: But that is your efforts to communicate to us about us reporting to you, rather than you doing a hygiene check to make sure that these things are not actually happening.

Rebecca Stimson: You mean something like proactively sweeping for content?

Chair: Yes.

Rebecca Stimson: Our AI is always proactively sweeping for things that violate our policies, with the varying degrees of success that I have described. There is no way for us to dial that up or down; it is as good as it is. What we can do is make sure that our ability to react to content when it is reported to us is even faster and better than it is normally during those times.

Chair: Thank you very much for your evidence today.