HOUSE OF COMMONS

Petitions Committee

Oral evidence: Online abuse and the experience of Disabled people, HC 759

Tuesday 19 Jun 2018

Ordered by the House of Commons to be published on 19 Jun 2018.

Watch the meeting

Members present: Helen Jones (Chair); Martyn Day; Luke Hall; Mike Hill; Catherine McKinnell; Paul Scully; Daniel Zeichner.

Questions 108-164

Witnesses

I: Karim Palant, UK Public Policy Manager, Facebook, Katie O'Donovan, Public Policy Manager, Google, and Nick Pickles, Head of Public Policy and Government, Twitter (via video link).

Examination of witnesses

Witnesses: Karim Palant, UK Public Policy Manager, Facebook, Katie O'Donovan, Public Policy Manager, Google, and Nick Pickles, Head of Public Policy and Government, Twitter (via video link).

Chair: May I welcome our witnesses, Karim Palant, who is from Facebook, Katie O'Donovan from Google, and, joining us on the video link, Nick Pickles from Twitter? Thank you for being with us this afternoon.

Before we start the questioning, I want to say a little word to Mr Palant, if I may. Your company has indulged in behaviour towards this Committee that we find totally unacceptable. On 15 May, Facebook agreed to give evidence on 5 June. Then with three days' notice, they informed us that they wanted to be here today with other companies, thus making sure that we couldn't have two split evidence sessions, as we had intended. When our Clerks said that that wasn't acceptable, we were suddenly told that the member of staff who was coming had gone on annual leave—something that, you will understand, we treat with some scepticism, given the circumstances.

I am also aware that your company's behaviour to other Select Committees of this House, in particular the Digital, Culture, Media and Sport Committee, has been less than helpful—taking a long time to send the Committee information that they requested, changing witnesses and so on. I want to make it very clear that your company will not be able to avoid democratic scrutiny; that it is not acceptable to try to disrupt a Committee inquiry; and that you do not dictate the terms of engagement—elected Members do.

If we feel that we do not have enough time to question you this afternoon, then we will recall you, and if necessary, we will issue a summons. You have given the impression that your company does not feel it has to be scrutinised and, frankly, that it has something to hide. In doing so, you have done them no service at all. Young men in your company may play games; you do not play games with the House of Commons. I hope I have made that entirely clear, and that we don't run up against this behaviour again.

I call on Paul Scully to open the questions.

Q108 **Paul Scully:** Nick, maybe I can start with you. We have heard that this issue has principally come out through disability, and we have heard a lot of evidence of widespread abuse of disabled people on social media. Why aren't you doing more to tackle this?

Nick Pickles: Good morning and thank you. Can I make a few abridged remarks to the Chair? We greatly appreciate the opportunity to do this by video link. We are in a strange staffing position of not having someone in the UK at present for a few weeks, so we do appreciate it.

I think it is fair to say that as an industry, we have stepped up our efforts on safety more broadly in recent years. Certainly, when I joined Twitter four years ago, the landscape was very different. The level of investment and the level of technology going into this has significantly improved. In reflecting on preparations for this Committee, the focus on types of hate crime, as opposed to hate crime focused on all people, has led to an uneven response.

I think that this hearing is highlighting an important area where we can do more, and where perhaps the response you see in other areas hasn't been mirrored fully in regards to this. Certainly we are looking to do more. You may be aware that in April this year, we changed our reporting function to explicitly call it out. Disability was covered under our behaviour and conduct policy. That was based on feedback from groups of disabled people. So there is more to do, and we are starting to make progress where we can.

Q109 **Paul Scully:** When disabled people were talking to us, they were obviously highlighting the fact that social media in general is a really important tool for them to connect. What sense of responsibility do you feel for making sure that they can use this important tool, but not get the abuse back?

Nick Pickles: No, absolutely. We would look at it as a failure on our part if someone's voice were silenced by abuse. We are very proud of the use that our platform is put to by people. Only yesterday, I was looking at a tweet that told me something I didn't know: if a service animal approaches somebody in the street, it's because the person that they work with has fallen down and needs assistance. Yesterday, that tweet, educating people about a scenario that might come up, had the best part of 200,000 retweets. That kind of expression is incredibly valuable to us, so if that is being silenced because of abuse, that is a failure on our part and a problem for us.

Jack Dorsey, our CEO, makes safety the company's No. 1 priority. It is something that we have been working on. We have made more than 30 changes in the past 18 months across our product, our policy, and our operations teams. We now take action on 10 times as many reports as we did this time last year, so we think we're making progress, but next month, in July, we will have our trust and safety council gathering in San Francisco, and "Who are the groups we need to hear more from?" is something that is on the agenda for that meeting. Actually, I'm planning on going to the trust and safety council and asking them, "How could we hear more from groups that work with disabilities?" because it's an audience that perhaps hasn't had the same level of engagement as other areas.

Q110 **Paul Scully:** Finally, what action would you take? You say you are taking action against more people. Can you give us some examples of what actual action you can take, apart from maybe barring their Twitter account?

Nick Pickles: When I joined the company, that was basically the choice we had. Four years ago, we could either suspend you permanently—kick you off—or we could give you a warning, and those were the only options we had. We now have a much more nuanced approach. At a basic level, we can ask you for a phone number, so if we think you're trying to use the platform and hide your identity, we can say, "Well, you can't tweet until you give us a mobile phone number and you verify that number is real."

Equally, we might say, "We agree you're engaged in passionate debates, but we're going to lock your account for 12 hours to give you a time out period." We might require you to delete a specific tweet that is in violation of our rules, but not so severely that we think you should be removed from the platform. We've introduced what some people have compared to a speed awareness course: a kind of step-by-step where you get told, "This tweet broke the rules. This is the rule it broke. Do you understand that? You'll now have a time out period, then you can return to Twitter."

I can send the Committee screenshots of that walkthrough. We have found that when people go through that alone, the overwhelming majority don't go through it again. When people get a warning, it does modify their behaviour. That's something where we need to think about interesting opportunities; there's a research project running now with MIT and—I think—Bath University, looking at how we can use reminders to users to improve their behaviour, rather than just kicking them straight off.

Q111 **Paul Scully:** Thank you, Nick. I will not go through everything with everybody, but Karim, could you say what responsibility Facebook believes it has to make sure that disabled people in particular can use it for the clear benefits that social media has, while cutting out this level of abuse that they are getting?

Karim Palant: Absolutely; I'd be very happy to. May I ask permission from the Chair to reply to the earlier comments?

Chair: No, answer the question that has been put to you, please. This is not up for debate; just answer the question.

Karim Palant: Okay. I will write to the Chair after the event with a correction to the record.

Regarding the question asked by Mr Scully, we take a great deal of responsibility. We believe that we have a very clear responsibility to our users and to those who use our platform to ensure that they can do so in a safe way, and for the content that people post on our platform. We feel that very strongly. We have taken a number of measures in the last few weeks to demonstrate some further commitment in that area, and I'd like to set those out in a little bit of detail, if that's okay.

The first thing is that we have long had very clear policies about abuse and hate speech on our platform and the abuse of individuals using our platform, and using language that may degrade, dehumanise or abuse individuals and attack them for what we describe as "protected categories." We have long included disability as one of our specific

protected categories within our community standards. I know that is a distinction in hate speech law in the UK.

In April this year, we published a much more detailed set of community standards. These explained in a great deal more detail not just the high-level standards that we expect our users to agree to, but the rationale behind them, and the guidance that we give to the people who review reports from users about that type of speech.

Every single piece of content on Facebook has a report button clearly on it, which allows any individual to report it to us. That is then reviewed against our terms of service, and it is removed if it violates those terms. As recently as last year, we became the first company to set out the number of people we have undertaking those reviews and reviewing that kind of content. It was increased to around 8,000 by the end of last year, and we have undertaken that by the end of this year, we will have 20,000 people working on the review and security and safety on Facebook—that is a doubling this year.

We also believe that technology can play a significant role in identifying this kind of abuse. It is not just about user reports; it is also about surfacing some of that content for review. We recently published a transparency report that sets out how much content we removed for hate speech in the first quarter of this year—that is the first time we have done this. It is an attempt to give people like you more of a sense and clarity about what it is we do. We removed about 2.5 million pieces of hate speech in the first quarter of this year.

We also set out how much of that was identified using artificial intelligence technology how much came from report. As you might imagine with something like hate speech, context and human understanding is key. Artificial intelligence finds about a third—30-something per cent.—of that content for services review, but the rest comes from humans through user reports.

Q112 **Paul Scully:** Roughly how much do you reckon the AI is getting it wrong, and eliminating things that are actually OK, because it has the context wrong and thought it was hate speech?

Karim Palant: This is one of the very big challenges. You need to set thresholds when you are setting these AI tools to work, for what level of accuracy you are prepared to accept. For anything below that, the people reviewing the content that is surfaced are spending far too much time viewing false positives, and not enough time reviewing accurately identified hate speech.

What we don't do at the moment is proactively remove a great deal of hate speech from Facebook, without it being reviewed by a human being afterwards. What we are doing effectively is surfacing content for review by human beings. We identify stuff that might well be hate speech, based on our algorithms, and then that is looked at.

There is one exception to that, which I think is of interest. It is a new step on Instagram. For some of the most egregious, really personally degrading comments attacking people's images—direct attacks on somebody's appearance, for example—we are starting to filter out the bullying comments, as we call them. That, by definition, is really challenging, as you will be aware. There may well be instances where legitimate speech is—

Chair: We will come back to AI in a minute. I will bring in Daniel Zeichner.

Q113 **Daniel Zeichner:** My question is also directed at Karim. Good afternoon. Terms and conditions are notoriously difficult to explain to people in general, and most people probably don't take much notice of them, but if people don't know about them, they are not likely to understand what kind of information might be being shared with others. Can you tell us anything you have done to ensure that your terms and conditions are accessible to people with disabilities?

Karim Palant: There is, I think, a distinction between terms and conditions and community standards; it is important to recognise that. The terms and conditions have to be a legally binding contract by law and there are certain things that bind on us as well.

Q114 **Daniel Zeichner:** Do you have an EasyRead version that people are likely to understand?

Karim Palant: Obviously, we have EasyRead and those kinds of tools available, but that is not sufficient in many cases to provide the kind of support that people with specific disabilities might need; they might need a more plain-language approach.

Q115 **Daniel Zeichner:** At the moment, is there something to help those people or not? Is that something you are looking at?

Karim Palant: In the context of online safety for disabled people more widely, we've been having conversations with a number of NGOs in the UK about trying to provide additional support. Many of the programmes that we run—for example, the digital safety ambassadors programme, which we run with the Diana award and Childnet—work with children with disabilities and particular vulnerabilities. They operate in those contexts, but we believe we can do more in that space to provide extra guidance and support for those young people and for people supporting them. It is something that we're actively looking to do.

Q116 **Daniel Zeichner:** So there is more to be done.

Karim Palant: Yes.

Q117 **Daniel Zeichner:** We have also heard from people who use social media to connect with others who are affected by disabilities. Sometimes they find out that photos of, and information about, their disabled children have been used to create abusive content. How do you explain to those users what happens when they share a photo or post information, and

how do you make them aware of the risks they are actually running? Do you have confidence that people in that situation are aware of what they are actually doing?

Karim Palant: There are a number of different ways to communicate that to young people, some of which are online. For example, when a young person sets up a Facebook account, the default setting is that it's only to be seen by their friends. We then provide a number of warnings to that young person at various points during the process. For example, when they accept a friend request, we make it clear to them that this person will be able to see their photos, content and so on.

When you post for the first time publicly, rather than just to friends—the audience is visible on every piece of content that you post—we will again prompt the young person at that point and say, "If you're changing your settings, this means that people who aren't your Facebook friends will be able to see this content." You can prompt within the app, but a lot of this is also about offline education and support. That is why, as of last year, we've worked with a number of different NGOs, including Childnet digital leaders and Diana award anti-bullying ambassadors, as part of our safety ambassador programme, to ensure that over the next couple years we're offering every secondary school in the UK a programme of digital safety—peer-to-peer education, where young people are supporting other young people.

Q118 **Daniel Zeichner:** Quite rightly, you're talking about what is happening with young people. I'm talking more about people with disabilities. Have you done anything specifically to warn them that they might be particularly vulnerable? We're not necessarily being critical—we're pointing this up as a potential issue.

Karim Palant: I've been reading through the evidence sessions for this inquiry, and in talking to NGOs that we work with regularly on safety—particularly around young people, women's issues and so on—there seems to be a general sense of this being an issue that we have all got a bit more to do on. I would agree with that.

Q119 **Chair:** May I ask our other witnesses to comment on that—on making terms and conditions easily accessible to people with disabilities, particularly learning difficulties? Katie, what is Google's policy on that?

Katie O'Donovan: For YouTube, which is probably the most relevant platform that we have in this space, we have a distinct, separate page that links our community guidelines. They're each captured in a bullet of very plain English, which we test and analyse to ensure that it is easily understood. We also have diagrams that represent each particular community guideline—they're not wholly descriptive, but they do help point you in the right direction.

More broadly, for all of the content we have online and the way we work with our different products, our core mission as a company is to make the world's information universally accessible. We absolutely review the accessibility of the content in the platforms we have. In those cases, it's

about ensuring that they work with screen readers, too. There is always an important amount of information to transmit, but we try to do that in very plain English and with diagrams. Most recently, we've also made a video that describes what happens to anyone who flags content—what happens on that journey—to help make that a little bit more accessible.

Q120 **Chair:** Have you particularly worked with organisations representing people with disabilities to develop those guidelines, especially with those who represent people with learning difficulties?

Katie O'Donovan: The guidelines in that particular content was created in the US, and I know they talked to a wide range of users to make sure that it worked. I wouldn't like to say that it absolutely included people with special educational needs, but I'm very happy to find out more and then come back on that.

Chair: Thank you. Nick, can I ask what your company has done in this direction to make the terms and conditions easily accessible?

Nick Pickles: Just to echo what Karim and Katie have said, the first thing that we did was to recognise that having everything in the terms and conditions is probably not the best way to go, and then separating things out that are relevant. I would call out two things. First would be the Twitter rules. Rather than having that as one long document, it is broken down into relevant areas, which you can navigate quite easily from our safety centre. Secondly, our privacy policy. Twitter is overwhelmingly public. I think that when people use Twitter they expect that it is visible to the world. That is one of the platform's unique points. Our privacy policy—similar to what Katie was talking about—has call outs, very specific highlighted areas and a way of drawing people's attention to not just the wording, but the practical implications. I can send the Committee examples of how we have done that. As Karim says, we have to have the legally required content. But also, how can we give people a pointer to say, "This is what it means, these are the practical implications or these are the changes you can make if you don't want this setting"? Those are all things we are looking at. One of the biggest challenges we have is that sometimes simplifying our policies makes them harder to understand. So there is a tension between adding more detail, so that people can understand, and making it simpler. That is something that we wrestle with across all of our changes.

Q121 **Chair:** All of you have talked about making things easy to read, but that is not the same as Easy Read for people with learning difficulties. Can I ask all of you to say briefly whether you have looked at having an Easy Read system?

Nick Pickles: I am happy to go first. I will come back to the Committee on that one. I am not aware that we have. I think the website and the way we deliver content is screen-reader accessible. So if people are using assistance devices to read the content, it is accessible in that format. But on the specific question of making this EasyRead, I think this is one of the

topics we could do more on, to understand where we are not doing things that would assist people.

Q122 **Chair:** Thank you. Katie or Karim, do you want to add anything to that?

Katie O'Donovan: We have taken action to try to ensure they are easy to understand, but I would like to look at it more to see if it meets the Easy Read criteria.

Karim Palant: I would echo that. We have done a lot of work, for example, on things such as social reporting. That was particularly done with young people and bullying in schools in mind, but it is used by adults and people of all ages as well. That was about providing a language that people can use to communicate with each other when they see content, for example, that they do not like. A lot of that has been done with the Yale Centre for Emotional Intelligence in the US, so it is very US-focused. It is about looking at language and the way people use it online with each other in an easy-to-understand way. I know that Easy Read is a very specific thing, so I will write to you about that.

Chair: Thank you. If you can send us some more information on that, it would be helpful.

Q123 **Catherine McKinnell:** We have heard from disabled people—as part of this inquiry—that they feel that the reporting of abuse should be more accessible to them, as well. One witness described how it would be helpful if there was just a big button to report abuse when they see it. What are you doing to make the reporting of abuse more accessible to disabled people? I put that question to you first, Karim.

Karim Palant: There is always a challenge to achieve a balance between the flow of users as they go through an app, and making everything really easy and accessible at the same time. But every piece of content, on the mobile app and on desktop, has a report button on it. Each photo or piece of text will have a report button there. There is a little thing in the corner you can click to report. As you will know, that is not always people's first instinct when they see a piece of content that they don't like. So that is obviously challenging. Where and how you respond to that report is also a really important part of it. We have something called the support inbox, which is in the app. That is a design decision that we took. Rather than emailing people off the app, we have a function within it whereby they will be updated on the progress on their report, and we come back and tell them what action we have taken on that.

I am very conscious that reporting and responding to those reports is only ever part of the solution, especially when somebody is very upset with behaviour that is often online and offline. We get a great deal of dissatisfied users and people who have been through difficult situations like that. I am not seeking to suggest that it is all fine, but as I say, we have a report button on each piece of content and we respond to every report.

Q124 **Catherine McKinnell:** Have you involved disabled people in the

designing of that system?

Karim Palant: We constantly refine how those buttons look, and we do user testing throughout. We try to track the flows of the people using those tools to ensure that they are being used as best as possible. We also have a wide range of NGO groups that we talk to, which have specific groups that they cater for. We test those flows with them and ask them for feedback and so on.

I have to say that, certainly in the UK, we haven't been as good as we could have been at dealing with disability NGOs specifically. A lot of the NGOs that we deal with will address and deal with disability issues, but it is not their main focus. We could do more work with those NGOs specifically to understand these issues a bit better.

Q125 **Catherine McKinnell:** Nick, coming to you, what accessibility issues does Twitter look at when designing its abuse reporting mechanism? You mentioned earlier that you had recently added disability to the list of options that people can use when they want to report abuse. Why has that only happened now and not sooner? What was the thinking in deciding that?

Nick Pickles: One thing it is important to say is that we always covered abuse against disabled people under our hateful conduct policy. The change was that, in the reporting flow, we listed some examples. It said "report hateful conduct" and then in brackets, "e.g. race, religion, gender". We had some feedback. For example, Muscular Dystrophy in the UK raised with us that that could create the perception that we didn't action abuse that was targeted at people based on a disability; simply by not listing it, it creates an expectation that the abuses that we listed are the things that we actioned.

We took that feedback on board and looked at a number of different options. One thing we are conscious of is that different people use different-sized fonts on their phones. If you add too much text in that box, it starts to run over several lines and becomes less easy to use. We tried to find that balance between how much extra information we can give without it making it visibly difficult to actually go through the reporting flow. We looked at a number of different options and we found that we could add disability. You will note that we haven't added every form of abuse that could be targeted at somebody. That was the thinking: we don't want to make it so complex that it makes it difficult.

When I joined the company it took roughly 15 clicks to report a piece of content. That is now down to about five. That was driven by the idea that, if you make something complex and have lots of steps, people who have special educational needs or are actually in a moment of need are less likely to go through the reporting flow. The challenge in making it simpler is that it also becomes more generalised. That is one reason we have a box at the end that asks if there is any context that users need to add that is particularly relevant. That is where that extra context around disability abuse, which might not be a violent threat, is really important.

Q126 **Catherine McKinnell:** Katie, do you think that the people who act when people report abuse receive enough training? Do you think there is enough understanding among social platform providers about the needs of disabled people? How regularly do Google provide that level of support?

Katie O'Donovan: All of our trainers who review flagged reports—we have made similar endeavours to the other companies to ensure that it is a very easy, quick process. You can flag content again, as you can with Facebook, from any piece of content, just by clicking on three or four clicks or, if you want to add a little bit more context, you do not need to add much information at all. Those reports—the reports of the trusted flaggers who we have recruited, who also work with different organisations and with different special interests to flag content that is very relevant—go to our teams of reviewers. Those reviewers are trained across our policies. In some cases, there are some with more specialisms or experiences, but we have a very clear community guideline that we do not allow content that incites hatred against disabled people, so as part of the training process that they go through, that is explained and they are given examples of what is and is not allowed on the platform. Before our reviewers go live to look at content and begin making judgments, they are tested on those criteria to ensure that they are making the correct judgments, as per our community guidelines.

We retest and work with people on a regular basis to ensure that, even after they have been testing for three or six months, they are still adhering to the community guidelines, rather than to any kind of unconscious bias that may have crept in or any change in perspective. We make sure very clearly that people are skilled in our community guidelines before they go live and on a regular basis. If we see any over-corrections or mistakes, we will retrain in those areas too.

One of the things that is also worth reflecting on is that we see different trends in what happens online. All of us have reflected that we have learnt things corporately as companies over the years or as individuals as we have seen different examples. Sometimes that can happen quite quickly with a trend or a phenomenon that comes online, which can be really positive or negative. In that case, we also ensure that our reviewers can learn and respond to something, even if it perhaps happened in one part of the world, by sharing it across the system.

Q127 **Mike Hill:** Following on from the theme of reporting, Karim just said, "We do respond to every report of abuse," and, very quickly, I would like to know whether that reflects your companies' performance, Katie and Nick. Supplementary to that, I would like to ask how long it takes to progress those reports, because we have heard from witnesses that it can be quite some time before they are progressed.

Katie O'Donovan: I will go first, if that is okay. Your first question was about responding to every report. We recently introduced what is called a user dashboard. Previously, if you made a report to YouTube, we received the report and we actioned it, but we did not actually communicate

anything back to you, so it was left to you to go back and see whether that content remained on the platform or had left. We heard feedback from many different groups saying, "Actually, when you are trying to build trust in the reporting flow and trying to encourage people to report content to YouTube, it is really important that you close the loop and let people know what happens to their reports." The user dashboard, which we launched about three months ago, does that.

When you are on YouTube, you can go into your account page and you will see a list of the content that you might have flagged. You will also see the action that has happened to that report. If you have flagged content—sometimes, we get flags that are inaccurate, erroneous or motivated by political or football rivalries—you might get a message that that content has remained. If the content did break our community guidelines, you will get a notification of what has happened to it.

In terms of how quickly we do that, the EU Commission is holding a review on how quickly all our platforms deal with hate crime. The most recent response is that they had found that in companies that were participating, 85% were taking action in 24 hours. For YouTube, 60% of action was taken in 24 hours, but that is absolutely our ambition and what our users should expect. The only caveat is that sometimes—this is more occasional than routine—cases are more complicated where the context is more important or where understanding the broader environment of which that content is part is important to us.

Q128 **Mike Hill:** The timeline is a bit more complex, but like Karim, are you saying that you deal with or chase up every report?

Katie O'Donovan: Yes.

Q129 **Mike Hill:** Is that the case with you, Nick?

Nick Pickles: Yes. A combination of humans or technology will review every report. Obviously, there are certain cases where, for example, duplicate reports are submitted and those kinds of situations. Similar to Katie, actually, one big action that we have taken in the past year is about telling people what we do. Previously, you would submit a report that would go into a black box, and you wouldn't hear back from us, and you wouldn't know what action had been taken. So now we are not only telling people the action we've taken but we're even telling them the account and the tweet that it is related to, so we're giving people much more context.

One thing that we're still working on is particularly relevant in this context, where we are talking about people who may have, say, special educational needs. It is how we can communicate with them to provide more context in a way that is helpful and accessible, because our platforms are working globally across different languages and different cultures, and making sure that we communicate in a way that is accessible is something that we still have a lot of work to do on. We have made progress in making it simpler and making it part of the reporting as well, but there are cases where we need to go back to the user and say, "We're not quite sure we understand

what's happened here. Can you provide more context?" That process can probably be improved.

Then the second part of it is on time. For us, we prioritise reports according to potential harms, so things like violent threats or reports of things that may involve someone at risk of suicide. Those reports will be prioritised above others, so you will see some differentiation between the type of incident and how quickly we respond to it.

Q130 **Mike Hill:** Both you and Katie have touched on the second question that I wanted to ask, which was all about reporting back to the people who themselves report the abuse, because some of our witnesses, including Katie Price, have told us they do not get feedback when they report abuse. Sometimes, as you said there Katie, they have had to go and find out themselves what had happened regarding abusive content and whether it has been removed.

I think you both spoke around what you are doing on that. What about Karim?

Karim Palant: Yes, absolutely. Anyone who reports abuse will get a notification in their support inbox that tells them what action we took. We will get to the vast majority of content within 24 hours. As Katie set out, there are complex cases sometimes that take longer, and we will prioritise, often using machine learning, against what we call real world harm. Clearly, with a threat of violence, we would triage that to be reviewed much more quickly than, say, spam or something else that doesn't necessarily have the same effect. So, we will notify people in their support inbox.

I think it's fair to say that there is a perception that often people don't get a reply and some of that may well be to do with accessing the support inbox and providing better education to help people understand that. We spend time on the platform and off platform educating people on reporting: how to do it; and where they will find the responses. But I accept that it is feedback that I think we all hear, and the industry as a whole hears, about people not feeling that they get a response. However, as far as we're concerned, it's almost an automated process, so we can be sure that in most cases people do get a response.

It's clearly also the case that often that response isn't satisfactory. Either the content itself is taken down but people—there is clearly a bigger issue there behind the report, which I think is what needs to be addressed. Alternatively, the content is not taken down and sometimes that's because it is very context-dependent and it is very hard for us to judge the context of a particular report. If, say, there is some bullying going on and a lot of it is going on offline, it's very, very difficult to be sure of that.

On the first point, where there is a broader issue of support, one of the things that we are doing at the moment is working—again, this is a focus with young people, but it is why there has been a lot of concern expressed, certainly in the UK—around where a young person reports bullying or abuse to us. They will be at the point of report. As well as us

looking at the report, they will be signposted to some extra support from ChildLine within the app, so we would direct them to ChildLine.

Chair: May I point out gently to you all that we are hearing a lot about young people—understandably—and special educational needs, which implies that we are simply talking about children here? May I just remind our witnesses gently that we are talking about adults, too, here?

Q131 **Mike Hill:** This is a final, quick question. Does anybody agree that given the context—we are talking about disability abuse—things could be improved in terms of accessibility of communication and feedback to individual complainants? We have even had witnesses say it would be nice to hear a human voice. Is that part of the offer that you have, Nick?

Nick Pickles: We do not currently offer that kind of being able to phone someone, for example. The question of how we support these users is something that we will take away. Certainly, it reflects the Chair's comment. Perhaps the groups that we speak to and the groups that are most vocal in Westminster, are groups that don't work specifically on these issues, so as a result we have over-indexed to other areas.

This inquiry is a welcome reminder that there are issues that have not received the same degree of attention, and how we can ensure that those users receive the support they need is something that we need to look at more, particularly beyond young people.

Katie O'Donovan: As I mentioned earlier, we do an accessibility review for all of our website, particularly on YouTube, including the report in flow. We have analysed it to ensure that it does meet the needs of disabled users, too. That includes having a very intuitive and easy-to-use reporting system.

We often find that people much prefer staying in the app or on the platform to do that, than sending an email or taking a phone call. We certainly take that feedback. To expand on Nick's point, what is always helpful for us is to hear directly from users, too. I have obviously looked at the previous evidence given to you by disability campaigners and disabled users and we will continue to analyse that but also welcome broader feedback on this issue.

Q132 **Mike Hill:** Thank you. Karim, have you anything to add?

Karim Palant: It is a really important point that it is not enough simply to have a reporting flow and to respond to reports. It is important to recognise what happens usually, in the vast majority of cases.

I was speaking this morning to some NGOs who work in this space, ahead of today's hearing. One point they made to me that really struck home is that a lot of what goes on—some of the abuse—is what some of this Committee have raised before, which is mate crime. At lot of that is people who are known to each other off-platform who are interacting on-platform. That mix of online and offline means that often support outwith the app is really important.

There is then a question. We should obviously have support available and work with NGOs and groups to help them provide that support. But are we the correct voice as a company to offer that support? Often not. That is why we do work with a wide range of partners, whether in the hate speech space around particular ethnic groups and so on, or whether it be disability or young people. That is the model that we have tended to pursue; to try to find partners who can offer that help and support.

Q133 **Martyn Day:** I am very pleased to hear that you all review the reports that are made, but if a report is made alleging disability abuse, how do your moderators decide whether it should be taken down? I will start with Katie.

Katie O'Donovan: Just to say that, on the reviews that we receive, as Nick mentioned, we use a combination of technology and humans to review those. Quite often you will get repetitive flags on particular bits of content, sometimes on the most egregious type of content.

Other times you might get a repetitive flag. One of the most flagged pieces of content we have on YouTube is a Justin Bieber music video. That is not because there is any hate speech in there or anything that breaks our community guidelines; it is just that people who are not Justin Bieber fans tend to flag that content. We absolutely use a combination of technology to help us prioritise, triage and identify the most relevant comments for our human reviewers to look at.

To answer your question specifically on how our reviewers understand and make that judgement call, we have our public community guidelines that start with a snapshot in plain English of what those community guidelines are, which prohibit in our case content that is inciting hatred against individuals or groups based on certain attributes, which include race, religion, disability, gender, age, veteran status and sexual orientation or gender identity.

We make that information publicly available and then we have resources internally for our community reviewers, so that they understand that in a little more detail. We give them practical examples of where something crosses the line and where something may be allowed. We are very proud on YouTube that we have a really vibrant and diverse group of disabled content creators, who create content both for other disability communities and their own disability community, but often reaching mainstream audiences and challenging some of the misconceptions or prejudice against disabled people. When we take those judgment calls we bear all of that in mind, but always return to our community guidelines, which give us that detail.

I also mentioned earlier that we work with trusted flaggers: organisations that know a particular area in more depth. For example, on content relating to suicide we work with the Samaritans. We also work with Stop Hate UK, which is an organisation that campaigns and takes action against hate crime of all natures in the UK. That again helps us. We find that all our trusted flaggers have a much higher accuracy rate than members of

the public who flag content, partly because they are less interested in Justin Bieber but very specifically because they have a great deal of information and an understanding of their own topic. They can understand in a very expert way what content needs to be flagged to us and why.

We work with trusted flaggers across a number of areas to help to inform our decisions. Sometimes those trusted flaggers will also identify emerging themes that perhaps we have not captured properly in our training, or that we have not been fully aware of and that we need to incorporate or include.

Q134 **Martyn Day:** Nick?

Nick Pickles: There are a couple of things that I can add to that, which I think were true for all the companies, but might be useful to call out. When a tweet is reported, we will look at the tweets around it to understand, for example, whether it is an argument between two people or one person targeting another, when the other person is not engaged. For example, if someone has been tweeted by name, or with their personal Twitter handle, repeatedly by the same person, that will appear different from two people having a robust argument where you all tweet. It is about the context around the tweets and the behaviour of those users.

I can share with the Committee that this is actually a policy area that we were conscious was one of hardest policies to understand in terms of when we will take action. We published a lot of information around the context of when we enforce to help people to understand. Broadly, the idea is that we do not want to stop people sharing opinions. We are looking at behaviour that is offensive. I think that is quite an important distinction.

Karim Palant: When a piece of content is reported to us and reviewed by one of our reviewers, that reviewer looks at the piece of content, looks at what context is around it, and will essentially compare it in a one-off judgment against our community standard. Essentially, the reviewer takes our community standard and says, "Well, this says you're not allowed to use dehumanising language to, for example, somebody who is from a protected category, whether that is race, religion or disability." That is essentially reviewed as a one-off against those rules, and is removed if it violates those terms. It does not matter if it is reported 10,000 times or once. It is reviewed in the same way.

It is really important that native speakers are doing that reviewing, especially on issues such as hate speech where context is very important. That means we have a lot of training around context in a particular market for people who are doing those reviews around hate speech. Nudity, for example, could be reviewed by somebody who may not have English as a first language because that does not, in a sense, matter so much, but when it is something that is around hate speech we obviously make sure that that is the case.

We provide extensive training to those reviewers. I mentioned earlier that in April we published a lot of the guidance that we provide to those reviewers. Previously, a lot of that had been kept private because we did not want to help those groups who might want to skirt close to the edge to do so. The judgment we took was that people such as yourselves, NGOs in this space, and people who might be victims of some of this speech may well want to look at our guidelines, comment on them, provide us with feedback, and tell us that they do not like them, that they do not draw the line in the right place, or that they are open to misinterpretation. We felt we were better off publishing a wider, more detailed explanation of what they were, to receive that feedback. As I said, we published the numbers of how much we took down in the first quarter of this year. It was about 2.5 million pieces of hate speech content specifically, which will include disability hate speech.

Q135 **Martyn Day:** You mentioned earlier some figures on the number of reviewers and security people you had. How many of them are based in the UK?

Karim Palant: We don't have any review teams based in the UK. A considerable number are based in Dublin in Ireland. They are obviously native English speakers and will have a good deal of knowledge on the UK market. We don't have review teams based in every country in the world, but we do ensure that we have a specialist team with familiarity with the UK market, and obviously a lot of them will be native English speakers.

Q136 **Martyn Day:** Obviously, the terms of abuse are always changing and there can be massive regional or national differences. Even within the UK, there will be regional differences. How do you ensure that your reviewers are kept up-to-date with those changes?

Karim Palant: It is constantly reviewed and we will constantly have teams of people internally who look at the reports that we are receiving to see if we are seeing any trends, but there is also constant contact with groups who work within the UK. It is a slightly different system than the trusted flagger system, but it is what we call effectively partners, whose role is to flag trends to us and to comment on areas where they think we are missing something—where either our policies are wrong or our enforcement is wrong. We are then able to have a dialogue with those groups. That is predominantly how we do it. It is a constantly evolving game and it is very challenging. It is challenging in this space; it is challenging in all areas as well.

Q137 **Martyn Day:** May I just ask the other members of the panel to tell us how many reviewers they have based in the UK?

Katie O'Donovan: The nearest review centre for the UK is our Dublin office too, which is our European headquarters. We have a team of reviewers there who I probably talk to two or three times during the week, if not on a daily basis. They have a really good understanding of UK culture and the UK context. They also have other resources available. As you mentioned, you can have regional dialect within the UK that can be very difficult to understand, not just on an issue like disability but on other

issues where we have had gangs and there is very local vernacular. Our teams make a very proactive effort to speak to people who understand the language or the context or the relevance, and also to use other resources to understand that too. They are not based in the UK, but they are excellent colleagues who work very hard to ensure that they do have the right criteria and the right understanding.

Nick Pickles: Our main local centre to the UK is Dublin, for Europe, Middle East and Africa.

One of the reasons why I relocated to San Francisco is to redouble the company's efforts, particularly with outreach. One of the things that I focus on, and the team that I am building, is around making sure that we are getting the insights from outside the company, from specialist groups, that perhaps either haven't dealt with us before or have dealt with us infrequently, to try to make sure that we get feedback more regularly and in more detail. Working with colleagues on product teams and policy teams, we can take that feedback and bring it into decision-making programmes at our corporate headquarters at San Francisco. That outside feed, as one of the routes we talked about into product and policy decisions, into headquarters is an incredible part of the process. That is something I am really looking forward to working on.

Karim Palant: May I quickly add something in response? As well as our review teams, a large part of our operation is around engineering specialisms and around developing machine learning tools and other safety tools that people may use. I just wanted to flag that we do have an extensive engineering operation here in London. We are actually this year doubling the number of engineers in London who are working specifically on these kinds of safety and user security issues. I wanted to give you that additional context.

Q138 **Martyn Day:** This inquiry came about as the result of a very specific petition, put forward to us by Katie Price, regarding the abuse that her son had received. Her son's name has itself been termed a term of abuse online. Would that type of disablist language fall foul of your existing policies?

Katie O'Donovan: It would very much depend on the context. You have a range of language that can be used, from words that we pretty universally all identify as being abusive and offensive and hateful against a protected criteria through to something where, if you and I were just discussing in the street without a back story or a criteria, they wouldn't even be abusive between us. We look at the context and the criteria and try and understand that. That is where, if you flag content, as well as saying "It's at this point in the video," you can add more detailed context to that.

We also empower our content creators to manage their own comments. So if you have a channel on YouTube, you can decide you want all comments on, everything goes and you want live, active discussion. In the vast

majority of cases, that is what our content creators choose to do and they engage in fruitful conversations. This goes through to you deciding that you don't want to have any comments at all, because the issue you are discussing or your own sense of enjoyment of YouTube doesn't rely on comments. In between those, we give the content creators different tools. For example, you can decide not to allow particular words in comments. I remember once talking to a beauty vlogger who had a small gap between her teeth. She said that quite often she received comments that she did not like and were hurtful to her, so she put a block on her comments to restrict that word coming through. Or you can hold all comments for review, because it's your platform.

Q139 **Martyn Day:** Obviously, if the term of abuse was a member of your family's name, you wouldn't want to block every post with them in.

Katie O'Donovan: No, but what you could do is put comments with that word in and hold them for review. If it is a big channel that is run on commercial terms, you can ask somebody in your team to review those to see which ones you want to go through and which ones you don't. You are able to do that on a personal basis, too. I understand that that isn't perfect and we are using machine learning to identify comments that incite hate against particular groups. That is becoming more and more effective and impactful on our platform.

Somebody's real-life name is exactly one of the most complicated areas that we would have because, absolutely, you wouldn't want to stop that word being part of the dialogue and the lexicon on YouTube, because there would be lots of valuable real-life use cases, in particular talking about somebody's son, where many comments will be positive and constructive. This is the kind of nuance in the tools that we hope will allow our reviewers to have the context and make decisions on that, while also empowering users so they can analyse what comments are coming in with that particular word and decide which ones they would like to be published and which ones they wouldn't. It is an excellent example of exactly why this is tough. It is not necessarily the word itself; it's the context that the people are using it in.

Martyn Day: Nick, can you add anything?

Nick Pickles: One of the things that happens—it's something we struggle with a lot—is that it's possible to be offensive without breaking our rules. One of the biggest challenges of offence is that it is subjective: different people are offended by different things. So trying to figure out how we enforce for that is a really big challenge, particularly where something has happened on national television and that content is then re-shared on social media. It is very difficult for us to action something that has been made very public, given that that content is in the public domain. Context matters, so things like posting to a user using @ handles versus being posted without to that user. This can be taken into account but this is also where the interaction with the law comes in. Again, something you'll see the CPS and the police struggle with is around where is the line between offensive speech and speech that shouldn't be allowed.

Karim Palant: I would echo the comments of Nick and Katie on the challenge of managing a term like that, which is somebody's name—a lot of people's name—and which could be used in that way. There are a great many challenging examples like that.

What I would say is that when it is your page and your Facebook presence or your Instagram, it is a lot easier to find solutions in those circumstances than it is on the open platform, where it is a lot more challenging. If it is your Facebook page, we can allow you potentially to block the name. As you point out, filtering out a specific word from comments on your page might not be the right solution in that situation, but you can block users from posting again and you can delete comments yourself. You do not need to go through reporting to us in order to do that—it is your page and you are in control. On Instagram, for example, it is possible to turn comments off completely, so if you are getting comments like that, you can do that.

Outside that, in the wider policy space and our policies in relation to particular words, we do look at trends, and we will look at the way people are using a word. If a word becomes used in a way in which it is clearly a term of abuse, we will try to train our teams to be able to spot that and to add that word to lists that they use in those circumstances. That is one of the reasons why an iterative, constant dialogue with groups that represent particular vulnerable groups is a really effective way of keeping track of those issues—but I will accept it is far from perfect.

Using technology and machine learning is capturing more and more of that kind of abuse, but again, it is very easy to see why that kind of abuse will be much harder for machine learning to pick up than anything else. Some of this is going to be about repeat actors being held to account. We will remove people's Facebook accounts if they are repeat offenders, and so on, so it will be about individuals being accountable for their actions as well.

Q140 **Martyn Day:** My final question is for you, Karim, in relation to Facebook. In your policy rationale for removing hate speech, there is a reference to "serious disability or disease". How do you define serious disability, and what does that mean about abusive content that might be about less serious disability?

Karim Palant: Disability as a whole is a protected category on our platform, so if you are attacking somebody on the basis of a disability, that is against our policies. I will look into the specifics of that line of the policy to see exactly what the detail is and write to you, if that is okay, because without having it in front of me I am slightly unsure of the context. There is clearly not intended to be a threshold—a level of disability—for how much we would consider something abusive. If you are abusing somebody for being part of a protected category, which includes disability, that is hate speech on our platform and we will remove it.

Q141 **Chair:** Can we go back to the use of artificial intelligence tools? Your recent report suggested that AI had flagged up about 38% of hate speech

posts. It is clearly difficult. How much is Facebook spending on developing that kind of technology, and roughly what percentage of your turnover is that?

Karim Palant: I'm afraid I do not have that figure. We do not ask our teams to account for their time in that way, but we do have a very, very significant investment in engineering, both in London and in San Francisco, which is focused on these issues. We are increasing the number of engineers based in London by 800 this year to about 2,300. A very significant proportion of those, both in San Francisco and here, are working on safety and security. As I said, we are increasing by 10,000 the number of people working on safety and security this year.

Q142 **Chair:** But you are not able to say how much goes into that particular area?

Karim Palant: I cannot tell you a specific accounting number for exactly how much. What I can say is that our chief executive and many other senior people in the company have talked about our increased investment in safety and security this year and said that we expect that directly to impact very clearly on our profitability in the short and medium term. I totally accept that there will be frustration that I cannot provide you with a full figure. However, I can tell you it is substantial enough that it will impact profitability for the company going forward.

Q143 **Chair:** What are the challenges and the limitations of developing that kind of technology?

Katie O'Donovan: We have touched on some of the limitations and some of challenges, but it is also worth reflecting on how the technology evolves over time. Certainly in the time that I have been working at Google and YouTube, which is three and a half years, the technology has greatly improved. In some cases that is because there is a greater amount of content to train machines on, to identify machine learning, and then, as Karim described earlier, to set a degree of tolerance for which you want the machine learning to identify content that then gets sent to manual reviewers. If you look at something like terrorism content, that has been a very successful way of dealing with that issue, partly because one of the strategies of those that upload terrorist content is to upload as much as possible, which is incredibly helpful when you are trying to train machines, because you need a high degree of reliability and a high volume of content to do that.

For something like the hate speech classifier, we have worked both on YouTube and in a part of Google called Jigsaw, which develops innovative technology for others to use. We have been successful in using machine learning to identify a classifier that can be used for comments. It is incredibly early days, but at the moment we are using that with a pleasing degree of success, particularly around abusive comments towards young people and children on YouTube, and we are looking to see how we can expand that. There are two points that I want to pull out of that.

We have the machine learning and the automated technology that we use that we have developed on YouTube for our own platform, but through the system that I mentioned at Jigsaw. They have developed a system called Perspective AI, which is open-source to publishers and to other forums and websites that have a discourse among members of the public, and they can use that to identify comments. It is not always that the machine will identify with 100% certainty, but what it can do is help prioritise what manual reviewers are able to do. That has been used by news publishers here and in the States.

Q144 **Chair:** Has that developed out of the technology you use to search for terrorist content, or is it a separate line of development?

Katie O'Donovan: I would not know with a degree of certainty—

Q145 **Chair:** If you could find out, that would be helpful.

Katie O'Donovan: You will find some characteristics are exactly the same and some are different, but the principles are the same and the technology will be more bespoke.

Q146 **Chair:** Thank you. Nick, would you like to comment on the challenges of developing that kind of technology and what Twitter has done?

Nick Pickles: This is an area of work where natural language processing, as it is sometimes known, is both very exciting in terms of the opportunities, but also still very challenging, particularly in certain contexts. Natural language processing does not really work for sarcasm right now; similarly, humour and posts that are made in a positive way that use certain words. The technology is not there yet to distinguish between those two types of comments.

To give you an idea of the scale we face, a 5% error rate on an algorithm could wrongly classify about 25 million tweets a day. One of the things that we are very concerned about is that if we use technology to take automated actions on accounts, even at very high accuracy, people whose accounts are wrongly actioned may be people who are using the platform to raise awareness about disabilities and to talk about their own experience of disability, so there is a very real risk that using technology that appears very accurate can still have a detrimental impact on the most important voices in this space.

We are certainly investing in this area. We acquired a London-based machine learning AI company called Magic Pony, and we are very excited about the potential, but the technology is not there yet. An academic at Lancaster called Claire Hardaker has done some really great work in looking at the potential for this, but in the short term the real risk that we are concerned about is that if you rely too much on technology that is not currently reliable, you risk silencing people who really need a voice, and that is a real cost. We talk about false positives and realising what that means: somebody who may be expressing their views about their own disability then loses that ability, and that is a real consequence that we are very concerned about.

Q147 **Chair:** Thank you. Karim, what difficulties have you faced in developing this technology?

Karim Palant: I would like to echo Nick's remark about the challenge of getting it wrong when you develop one of these classifiers that can remove content based on the best understanding that it may well violate your policy in a certain way. One of the real challenges in the counter-speech space, where you have NGOs and groups that are trying to counter some of the negative narratives by promoting positive speech—whether on hate speech or counter-extremism, or positive ideas of disability—is that those are the groups that feel it first. We feel that very directly, with them contacting us to say, "My page has been taken down," or, "An advert has been removed." That is often because it is not necessarily about AI but because a lot of the content that they are posting may well look quite similar to content that is violating policy. That is one of the biggest challenges—avoiding false positives, and getting it right.

Another challenge, which I think Katie alluded to, is about finding datasets that are reliable—the correct dataset to teach the algorithm correctly what is and is not violating content. We need a certain number of reports that pertain to a certain type of content and that are then correctly reviewed by human beings, who will say, "Actually, that's violating hate speech against a disabled person," or, "That isn't," and essentially build up a dataset that it can learn from. That is one of the big challenges. Once you have built something like that, you will surface a lot of material, some of which will be violating of our policies and some of which will not. Then you obviously need a lot of human resource to review that content, continuing to create that feedback mechanism. It is not a quick and cheap fix. That is one of the real things it is important to lay out.

Q148 **Chair:** Katie, in 2017 the Home Affairs Committee found that Google was using its technology to identify illegal and extreme content to help advertisers. Have you changed your practice since that report?

Katie O'Donovan: I am not quite sure I follow the question. We were using the technology to alert advertisers to terrorist content—

Chair: The phrased used was "illegal and extreme content".

Katie O'Donovan: We would remove any illegal or extreme content that we identified on the platform, because it would break our terms and conditions, and our community guidelines.

Q149 **Chair:** May I read you what the Committee said, to be helpful? "Google is currently only using its technology to identify illegal or extreme content in order to help advertisers, rather than to help it remove illegal content proactively. We recommend that they use their existing technology to help them abide by the law and meet their community standards." Have you taken what that report says on board?

Katie O'Donovan: That was the early 2017 hearing, rather than later ones. So yes we have, absolutely. That is partly because of the improvements in technology and in the accuracy. We have obviously

engaged with a number of Committees of the House and in other countries, and heard clearly from stakeholders in law enforcement and government about the action that they have taken and would like us to take. We have also been able to develop the technology to identify that. So yes, we use technology to identify proactively in particular areas where we have very good dataset of content that we think might be in violation of our policies, and to review it and review it quickly. The area we have made most success of terrorist content. We did take that recommendation on board.

Q150 **Chair:** Thank you. Nick, I understand that Twitter can lock an account, and that it will not unlock it until an offensive tweet is removed. Will you give us any idea of how often you have done this? What percentage of offensive posts have been dealt with in that way?

Nick Pickles: We have not yet published that data. It is something we are working towards. As Google and Facebook both mentioned, they have recently published expanded transparency reports. Our transparency report does not currently cover the full range of actions that we take, but it is something we are actively working on. It would be useful information. The one piece of information that I can share with the Committee is that the number of accounts that have gone through that locking period, and the number of them that don't go through the locking period again, is roughly in the mid-60s per cent.—I think it is 67%, but I shall double-check the figure. One of the challenges for us is that it is one thing to publish the raw data, but helping people to understand what the data mean is an important part. We are working on that now. I am hoping before the end of the year that will address part of it.

Q151 **Paul Scully:** First, I have a brief question for everyone. Germany is far more stringent on anti-hate speech, especially in recent months. Whether it is human responses or machine learning, do you deal with different languages differently?

Karim Palant: Obviously, one of the challenges of running a global platform is the vast range of languages you can be dealing with. Our community standards apply in whatever language you are posting in. Our goal is to have native speakers reviewing reports in the native language for any language. That is true in Germany and in the UK. For our AI classifiers, often we will have far more data in English and it will be easier to develop that.

Q152 **Paul Scully:** Would you apply the same set of rules?

Karim Palant: Absolutely. The community standards are the same. Clearly there are words and particular slurs that will vary in the team that is focused on each market. That is not necessarily just a language thing—a slur may mean different things in different parts of the UK, never mind between the UK and Ireland and the US and so on.

Nick Pickles: One thing I would flag is that, as Karim says, our rules are global; there are contexts where we would take action on an account in one area and we would withhold it from that country. The obvious example

might be that in Germany, even prior to the new law, there were very specific rules and laws about the use of Nazi imagery. That would mean that imagery that we would see in national newspapers in the UK would not be allowed in Germany, so we would geo-block that. That is a good example of where countries have specific and clear laws about what is and is not allowed, as opposed to laws that perhaps use terms such as "offensive," which are much broader. There were situations in which we would take action in Germany but not in the UK, but the vast majority of content that would break any country's laws is likely to be against our terms of service.

Katie O'Donovan: That is exactly the point I was going to make. We have universal community guidelines that operate in any market, but of course different countries have different laws and we abide by the laws in the countries we operate in. As Nick alluded to, Germany has stricter rules on Nazi iconography, and holocaust denial in Germany is illegal. We may take additional action in countries where there is an additional requirement as per the law.

Q153 **Paul Scully:** How do you make sure that your staff have that comprehensive knowledge of the different jurisdictions?

Katie O'Donovan: We do that through training. For example, in the UK we have a list of proscribed organisations under UK law and for Northern Ireland. Our staff around the world have very thorough training on what those proscribed organisations are, what the emblems of the organisation might be and their names or any other names that they are known by, so that they can make that decision based on the law of the land.

It is not always easy; it was mentioned that sometimes we will make mistakes in one direction or another. Sometimes we overcorrect because we want to make sure that no content that seeks to recruit people to terrorist organisations is on the platform. Occasionally we make an overreach. A recent example was when we took down content that had been uploaded by an important citizen journalist in Syria because we had misclassified it as terrorist content.

You will find different examples in different areas, where there is a live platform with very complex context, an international set of rules, a local set of rules and our own community guidelines. We are investing heavily in having the people who can review the content manually, because they understand the context in a way that machines certainly do not at the moment, but also in fully training in the areas that they are leading reviews of.

Q154 **Paul Scully:** Nick, you picked up on the word "offensive" and rightly said that it can mean different things to different people. In your view, is the current UK law fit for purpose to tackle online disability hate crime?

Nick Pickles: Yes, I noted that in the petition itself there was a call for a new offence for online abuse. I think the work that the Law Commission is doing is incredibly important here—we have certainly met with them. I think that sometimes the phrase "What is illegal offline is illegal online"

gives a degree of confidence that the law is very clear. Actually, when you have offences such as section 127 of the Communications Act 2003, which is specifically applied to electronically transmitted messages, I think there is more that can be done to improve the clarity of the law.

You might recall a trial that became known as the Twitter joke trial, which went to a very senior court, the Court of Appeal, when it was thought correct to prosecute, and ultimately it transpired that this was a badly made joke that got to a very senior court—so I think it is about both the provisions that are in the law and the guidance that goes to police and prosecutors. Something that I have become concerned about in the time I have worked in this space is that for repeat offenders whose cases perhaps should be dealt with as harassment cases are dealt perhaps as more straightforward cases, where a single tweet is used to prosecute rather than a pattern of behaviour. So I think it will help to clarify, but it will also help the police because for them it will be much clearer where the line is.

Q155 **Paul Scully:** Thank you, Nick. Karim, the petition calls for it to be a specific crime, as we have heard. What is your view of that suggestion, and how would it affect your way of working?

Karim Palant: I think that to some degree this is more a matter for Parliament than for us as private companies, but I will say that we find it helpful to have disability as a protected category in our rules. That is very clear and we have clearly set out what we deem to be hate speech against a protected category on our platform. We find that very helpful, in terms of being operable at scale. Our experience of working with law enforcement and prosecutors in the UK is that there is this sense that potentially a great deal of speech could be illegal under some definitions, but it is unclear in a great many cases and there is a sense that greater clarity would help both law enforcement and companies such as ours to respond to specific cases.

Another important point here, as Nick said, is that some of this is about behaviour, rather than specific phrases and words. Some of it is about repeat actors—so people who may well say something that you could define as a crime; it may be quite a high bar to define it. You could define it as against our terms and, again, it could be quite difficult. But if you look at a pattern of behaviour over time, actually they are engaging in some quite unpleasant behaviour, and finding a way to capture that is also really important.

It is not necessarily about specific words or phrases; it is actually about behaviours. When somebody is prosecuted for an offence, that might be one tweet, one Facebook post and so on, but actually if you can capture their wider behaviour in an order, or something like that, that would give greater clarity, because then they can't just move from platform to platform repeating that behaviour. There is a clear legal framework. Due process has taken place and they have been found to have behaved in a particular way that means they shouldn't be allowed to continue to do so.

Q156 **Paul Scully:** So there is a wider issue, isn't there? You talk about working with law enforcement, and about using and sharing illegal content in general. The Home Affairs Committee heard evidence from the police—again in February 2017—that the social media companies were not doing enough to remove illegal content. Since that time what have you done to improve your collaboration with the police?

Karim Palant: We have been working closely with the Met police's online hate crime hub over the last few months—the last 18 months—and we have been working closely with the Home Office and law enforcement on the removal of terrorist content. Our transparency report, which came out a few weeks ago, shows some real progress in the removal of terrorist content: removing 1.9 million pieces of terrorist content in the first quarter of this year, which is sharply up on the last quarter of the year before. So we are making real progress. I think we could be doing more intelligence sharing with local organisations on what trends they are seeing, and how we can proactively get ahead of some of those.

Q157 **Paul Scully:** Is that something you are now going to do?

Karim Palant: It is something that we are already doing, and will continue to expand.

Q158 **Paul Scully:** If you came back in a year's time, would you be able to say, "We've done this, this and this"?

Karim Palant: Yes, absolutely.

Q159 **Paul Scully:** Fantastic. Nick, same question to you about working with the police and greater collaboration since that report.

Nick Pickles: Karim touched on City Hall and the Mayor of London's online hate crime hub. That is really important to call out, because it was not just the police and industry; civil society was in the room as well. They were able to assist both the police and industry with context. That was primarily through Stop Hate UK. That was an excellent model. We have certainly been made aware of the Home Office's plan to have a national online hate crime hub, and we are awaiting further details on how that will work. Again, I think the collaboration between charities that work on abuse and hate crime, the industry and the police is very important.

I testified to the hearing in February 2017. One of the points that I raised then was that it is not our job to decide what is illegal; our job is to decide what breaks our rules. It is time to start challenging the idea that something is illegal so we should remove it, without due process having ever decided that something broke the law. It is our job to review against our terms of service and our rules. It is not our job to review what is illegal; that is up to the police and the courts. There is an important distinction to draw there. On Karim's point, if you take terrorist content, fewer than 1% of the accounts we remove were notified to us from a reference. The vast majority of this work is done through our own proactive features and user reports and working with NGOs. That is the area with the greatest focus.

I have also spoken to the College of Policing about the training we can do, making sure that police officers know how to request information from us, and the legal process to assist with investigations. We have a team that does the training with police officers. It is also about making sure that if a victim goes to the police force, the person on the front desk can converse with them about what has happened. In the case of violent threats, for example, we launched a plan essentially offering you the choice of getting an email setting out everything that you would need to go to the police—the relevant links, the time, the days, the accounts—so the person behind the desk has everything they need. There is still a bit of a skills gap in some parts of law enforcement, particularly among enforcers who might not deal with digital crime very regularly.

Q160 **Paul Scully:** Brilliant. Thanks. Finally, Katie, Katie Price's petition also calls for a register of offenders. Do you think that would help you to tackle abuse at all?

Katie O'Donovan: We look very carefully at content on our platforms that breaks our community guidelines. We work with the police and other community organisations as well. The idea of registers for offenders for any type of crime is sometimes controversial in the UK. That is something that would take a lot of parliamentary scrutiny. I am not sure it would always help us because we operate global platforms. If you just had a UK-only register of abuse, it would not be relevant if we had comments or content that broke our guidelines from France, Ireland or elsewhere.

What is important is that we have on our platforms measures to tackle repeat offenders. If you upload content that breaks our terms and conditions or our community guidelines, we will strike your account against each piece of content that we remove. That is the same for comments. We will take action against persistent offenders, including removing their accounts. If you have spent a lot of time developing a YouTube channel with a number of subscribers and followers, that is a very severe penalty. People are aware of that, so it drives positive behaviour and adherence to our community guidelines.

Chair: I want to wrap up before the vote, but we must get Catherine McKinnell in to ask about mate crime.

Q161 **Catherine McKinnell:** Yes. I know you have mentioned it already, Karim. The Chair was absolutely right to highlight that this is not just about children, but about adults with learning disabilities. We have taken evidence about the fact that there are disabled people who have been befriended with the intention of exploiting them, and some of that takes place on social media platforms, particularly Facebook. What is Facebook doing to tackle mate crime on social media platforms?

Karim Palant: Thank you very much for the question. It is inherently a very challenging problem to solve using tools or technology, because to all intents and purposes it looks, certainly in its early stages, like a genuine friendship and a genuine attempt to reach out to somebody and make

friends with them. It is incredibly difficult to provide a technological solution for that.

I have spoken to a lot of NGOs in this space and read through some of the evidence for this session, and a lot of their feedback is about the fact that a lot of these are pre-existing relationships that started offline and are being carried on online. That is part of where it develops. It is complex, because it is often across several platforms. It may be on a dating platform or a gaming platform, and then it comes over to Facebook. That is by way of context for how big a challenge this is.

We talked earlier in the session about how we can make the reporting functions easier to access for people with disabilities, who are particularly vulnerable to these issues. We want to talk to the NGOs I was having a conversation with earlier about that. I certainly want to do that, and there is more we can do to make that clearer.

Some of this will be about offline support, education and the tips and guidance that we can provide. We need to work with groups that offer support to those individuals in their offline world, and help them with tips, support and advice. As I say, it will be incredibly challenging to find a technological solution whereby we identify the individuals who are at risk or are engaged in it. We are engaged in dialogue about that, and we want to continue that with the groups that are particularly focused on this issue, because it is very complex and challenging.

Q162 **Catherine McKinnell:** It is reassuring that Facebook is engaged in dialogue. I completely agree that these things sometimes start offline and move online, but the really harrowing cases start to emerge when it moves from online to offline. I don't get the impression that that is a high priority. I appreciate the difficulties that you describe, but the purpose of this Committee is to ensure that it becomes a high priority for social media providers. One of the big issues about reporting is, as you say, that a lot of the people who are being exploited on these social media platforms do not realise that they are being exploited. Raising awareness is a big issue. Is there more you can say about how Facebook may be able to help raise awareness of these issues?

I have one other question. In the last session we picked up on the difference between your community standards and enforceability, in terms of the law. Would it make it easier for you to be able to remove this type of behaviour and abuse if abuse of disabled people was a specific offence—at the moment, it is an aggravating feature of a criminal offence—or does that not impact on your community standards? Does that move into the enforcement side? I would be interested to hear your comments on the first question, Karim, and those of the other panellists if they have anything else to add.

Karim Palant: Raising awareness and helping the people who are at risk and those who know them to understand the risks is a really important part of trying to tackle this problem. It is fair to say that, while this issue is really important, this Committee and these hearings have highlighted it,

which means that those conversations will happen with a great deal more focus. That is a very welcome outcome of these Committee sessions.

On the specifics of the law, as I understand it a lot of this is already a crime. The clarity around the law would help in some circumstances, but in the vast majority of cases, when we are talking about removing content, our community standards will already cover a great deal of that. Clarity in the law would help in a great deal of circumstances, and may help some of the other organisations that support those individuals, quite possibly, but in terms of removal it wouldn't make a huge difference. Where it might make a difference is if there were individuals who might be prosecuted for behaviour who might not otherwise be, and who may therefore have orders placed on them that might well prevent them from repeating that. That could be something that might be looked at as an area where you could find real progress.

Q163 **Catherine McKinnell:** Do any of the other panellists have anything to add about clarity in the law, or any other issues? These are our last questions, so is there anything that you think would be helpful that we haven't heard already?

Katie O'Donovan: In terms of clarity on the law, through our community guidelines for us and on YouTube we have created the clarity we need to be able to act on this content. We are investing in the technology and people to enforce our community guidelines. I don't think a change in the law would necessarily be helpful in that very narrow application, although I think that we have heard very compelling evidence from others who feel that it is important. One benefit it might have is reminding people that this is unacceptable behaviour. There is a societal benefit and responsibility for all of us to take action, whether that is seeing and reporting, or acting on the report of such content.

Nick mentioned the ongoing Law Commission review into online crime and law. I know that they have similarly done a review previously on disability law. They will obviously have some very important outputs on that, but it doesn't restrict or impede us from taking action. I think it is really important that we still have the flexibility to take action against offensive and hateful content.

Nick Pickles: I would just echo Katie's point. What worries me about this space is that sometimes the solution is perceived as removing content and accounts, and the part about holding offenders to account and ensuring real-world sanctions is sometimes lost. So I think particularly in these cases, which are particularly broad, that is where the real-world sanction has to come in. If there is more that we can do as an industry to work with the police so that they have the information they need to prosecute these people, that should help. There is definitely work we can do with charities to raise awareness. I think some of the best uses of our platforms have been able to raise awareness of issues, which is really important.

One point that I think is very important is that there is an opportunity to use sentencing orders so that where someone is convicted of a crime,

there is then a prohibition on them operating on social media. They are required to remove accounts, but then not to return on platforms. That is the kind of evidence we can act on very quickly. That means that the due process is proportionate and allows for appeal, but once that due process is concluded, we can act on offenders. In the case of hate crime, one of my big fears is that these offenders might be determined enough that they will use different platforms to target different individuals, and we need to make sure that we take a robust, criminal sanction against them to deter that behaviour, as well as just removing it from social media.

Q164 **Chair:** I thank our three panellists. You have given us a great deal to think about in the preparation of our report. I particularly thank Nick for getting up early in San Francisco. We are grateful to you. As always, if there is any further information that you want to give us, which you feel you haven't had the opportunity to say this afternoon, please feel free to write to the Committee and send that information. Thank you very much indeed.