HOUSE OF LORDS

Select Committee on Communications and Digital

Corrected oral evidence: Freedom of expression online

Tuesday 27 April 2021

3 pm

Watch the meeting

Members present: Lord Gilbert of Panteg (The Chair); Baroness Bull; Baroness Buscombe; Viscount Colville of Culross; Baroness Featherstone; Baroness Grender; Lord Griffiths of Burry Port; Lord Lipsey; Baroness Rebuck; Lord Stevenson of Balmacara; Lord Vaizey of Didcot; The Lord Bishop of Worcester.

Evidence Session No. 26 Virtual Proceeding Questions 207 - 213

Witnesses

I: Katy Minshall, Head of UK Public Policy, Twitter; Richard Earley, UK Public Policy Manager, Facebook.

USE OF THE TRANSCRIPT

Examination of Witnesses

Katy Minshall and Richard Earley.

The Chair: Our witnesses today are Katy Minshall, head of public policy for Twitter UK, and Richard Earley, Facebook UK's public policy manager. Welcome, both of you. Thank you very much for joining us. Today's evidence session will be broadcast online and a transcript will be taken. I think you have been following our inquiry into freedom of expression online, with its strong focus on the role of platforms, the harms caused to users and the response of the platforms to protect users from harm while providing for freedom of expression online. We have a number of issues we want to talk to you about today. If you have any brief words of general introduction, please fire away in answer to the first question, but let us get stuck into the questions.

Q207 **Baroness Bull:** Many thanks to you both for being with us today. Throughout this inquiry, our discussions have ranged from the technical and the legal to the various factors that act on, encourage or even discourage what we might describe as good, civil behaviour. We have talked about the role of education and the role of design. In a previous report, the committee suggested, "Some internet technology is deliberately designed to take advantage of psychological insights to manipulate user behaviour". That is a quote from our own report.

We have also heard from a range of witnesses that business models and, indeed, success as a platform user are dependent on less than civil behaviour. You attract followers by being forceful, provocative, even rude. We have heard also that the need to clock up likes and followers feeds addictive behaviours. How can design features on platforms better encourage, or even reward, civil behaviour, or is civil behaviour in fact antithetical to your business models?

Katy Minshall: Thank you for having me and Twitter here today. Safety by design should be the north star for all tech companies. The step change I have seen at Twitter over the past couple of years has been exactly as you said: how can we nudge people towards the behaviours we all want to see online?

I will give you an example of how we have gone about that. One of the behaviours we want to encourage is critical thinking: people thinking before they tweet or share something. To incentivise that, last year, we started experimenting with a nudge in the platform. If you go to retweet an article, we will ask you if you want to read it first.

The results we have seen are really promising. We saw a 33% increase in people reading the article and a 50% decrease in people subsequently sharing it. That is an example of where, by building a nudge or some friction into the system, you can, in quite a tactical way, start to encourage the behaviours you want to see online.

Baroness Bull: Can I push you on that? Does that encourage civil behaviour? You talked about safety, rather than civility. The two are connected, but I wonder if you can drive a wedge between them for me

and think about this question of civility.

Katy Minshall: It is a great question. On the one hand, we are now deploying this kind of thinking to the challenges of how people communicate with each other online. At an earlier stage, we are trialling a similar prompt. If you tweet something out, I go to reply to you and what I want to say looks like it might be harmful, be toxic or break the rules, I might see a prompt saying, "Do you want to revise your reply? It looks like it might be harmful. Are you sure you want to send it?" It is at an earlier stage, but initial results are promising. It is an example of how you can apply that thinking to issues such as civility online.

Twitter gives everyone the right to speak publicly. That has been a huge societal shift over the past 10 to 15 years. We have seen many people who traditionally lacked equal access to public platforms, particularly young people and members of marginalised communities, speaking up on Twitter. Twitter reflects a mirror on society, but that does not mean that Twitter should always be a wholly positive or nice place. Twitter at its best has sometimes been the exact opposite of that, with conversations of huge importance, like MeToo and Black Lives Matter.

By the same token, those discussions also highlight exactly what you are saying. You should always feel safe expressing your unique point of view online. It is our responsibility to make that happen.

Baroness Bull: I have one more question, about the business model. What feedback are you getting? You have talked about a reduction in activity in response to these nudges. Are you conducting any research to find out whether people actually like it?

Katy Minshall: That is a good question. The outcome we are looking for is that people are mindful about what they are tweeting. Concurrent to that, we are interested in whether that is something people want more or less of. Critically, with the harmful replies prompt and the retweet prompt, the main question we are trying to ask is whether this incentivises the behaviour we want to see.

Baroness Bull: Richard, can I turn to you with the same set of questions about how design can encourage good behaviour? How much is that in tension with the business model?

Richard Earley: Thank you very much to the committee for inviting me to speak today. I am really glad that you asked this question about the link between extreme or sensationalist content and platforms' business models. There is this idea out there that it is somehow in the interests, commercially or otherwise, of platforms such as Facebook to promote this kind of content. That is categorically not the case. It is not something that our users want to see. It is not something that the advertisers, which we use to pay for our service, want to see. It is not in the long-term interests of companies such as Facebook either.

We know that users do not want to see hateful, false or divisive content, because they tell us that all the time. We survey our users constantly and they tell us, and we know, that they do not want to see this. We carried out some research in Germany, which we published two months

ago, that showed that more than 50% of people are less likely to comment or engage in a conversation when they see hate speech online. Similarly, more than a third of people know somebody who has stepped back from using social media because of those kinds of conversations.

On the advertisers' side, Facebook knows this very well. Last year we saw a very large, co-ordinated boycott of our services from advertisers, because they do not want their brands or products to appear next to divisive and hateful content. As a platform, if we were to build our systems in a way that encouraged people to spend 10 or 20 minutes longer on our services, that might be good for our bottom line in the short term. In the long run, if people leave our services with a negative impression of Facebook, they are less likely to want to use it or any other services that we develop in future. That is why, over the past several years, we have developed a number of tools and made a number of changes to our systems to discourage this kind of divisive rhetoric. There are the policies we have and the tools we use to remove harmful content, which I can talk about a bit later.

On your question about civility, we have done a lot to tinker with our algorithms to make sure that, rather than promoting divisive content, they are used to reduce it. In our newsfeed, which is the thing you see when you first open up Facebook, we use technology to show lower down the kinds of content that people tell us they do not want to see, such as false content, clickbait and engagement with debates.

In listening to that, we have increasingly built tools for people to use to have more control over the kinds of posts they see in their feed. In Instagram, for example, we have rolled out a feature called "muted words", which means that you do not need to see comments on your pictures that contain words or phrases that you find unpleasant. Just last week, we expanded this to people's instant messages too. There is a lot we do to remove bad content. There is also a lot we do to try to encourage civility. We are thinking a lot about what more we can do to promote and encourage a more positive dialogue on our platform.

Baroness Bull: I think we will come on to it in the next question. I do not want to hog the floor, but you said that, if people see hateful content, it encourages them to step away. Would it not be better if there was a way of encouraging them to step up when they see hateful content?

Richard Earley: That is an excellent point. We have quite a number of partnerships around the world with what are called counter-speech organisations: organisations that are looking to encourage people to engage and challenge hateful narratives. There is a lot of great work being done around the world on this, but our role as a platform is to support those who are engaged in those kinds of activities, while making sure that our users have as much control as possible over their experience.

On the civility piece, I said that we are always surveying our users on their thoughts on this. In fact, just last week we announced a new set of global tests that we will be running, asking a very large percentage of

our users to give us feedback on the posts they see on their newsfeed and to answer new questions. Do they find them inspirational? How relevant are they to them? By looking at posts that drive more angry reactions, as we call them—you can click on a post on Facebook and like it, heart it or put an angry face—we will be doing some tests on that to see what the impacts of deprioritising that kind of negative content would be.

This builds on a commitment that we made in February of this year, when we said we would start to reduce the overall proportion of people's newsfeeds that contained political content. That was due to feedback from our users, again, that they did not want political content to take over their platforms. It is already a very small percentage of what people see on Facebook. Only around 6% of people's newsfeed is political, in the US at least, but we know that people want to see more from family and friends. We want to provide them with that.

Q208 **Lord Vaizey of Didcot:** You have covered quite a lot of ground already, but I wanted to slightly flip it. You have talked to a certain extent about what you do in the back office, as it were, in reducing content, making sure that you can take it down and so on. I will not disappear down that rabbit hole. What are the sorts of things you can do to help me, as a user on the interface, reduce my exposure to harmful content, without having to go to the trouble of taking it down?

We had a discussion with a previous witness, last week or the week before, where I thought they made quite a helpful point. In this debate, we are, quite rightly, holding the platforms to account, to a certain extent. There is also the possibility that, just as you might choose to walk down a brightly lit road rather than through a very dark park, if you are coming home late at night, users can use tools themselves. They can block. They can mute. They can unfriend. What kind of room is there to give users more of those kinds of tools?

Richard Earley: You are absolutely right. There is a lot that platforms can do to help users have that control. I spoke a bit about it just now. We have our policies on Facebook, which are called our community standards. They are the rules we write with experts all around the world for the types of posts and images that are not allowed on our services. They cover things such as hate speech, bullying, harassment and incitement to violence.

Then there is a big chunk of posts and speech on our platform that, while not breaking our rules, some people might find objectionable or offensive. A lot of the work we have done has been trying to give our users the ability to create their own spaces, where they can see things they want to see and protect themselves from posts they do not want to see. A very good example of this, which I just talked about, is the muted words or hidden words feature, where you can prevent certain comments appearing on your Instagram posts. Something similar that we launched two or three weeks ago is greater control over who can comment on posts that you produce, whether you are an individual user or a page.

It has always been the case, as you rightly said, that you can choose to unfriend people. You can choose what privacy level your posts will have. We have now gone down another level of granularity and given people the ability to control who can comment on posts they are putting on Facebook. We think this is very helpful for public figures, including Members of Parliament or the House of Lords, who want to make sure that they can have a conversation with their constituents in a way that is meaningful and safe for them. There are more tools like that that we can use.

We have, as I said, introduced some additional controls on what people see when they open their newsfeeds. I mentioned that newsfeed is the main part of Facebook that you open up. There have been a lot of questions, understandably, about how the pieces of content that you see in the newsfeed are ordered by our algorithms. We want to give people more transparency on that by publishing how these things work, but we also want to give them more control. We now have the ability for someone to switch their newsfeed to a purely chronological feed if they want to, without any algorithmic sorting, or to choose to see things from their favourite people first.

Lord Vaizey of Didcot: How easy is it to use this stuff? I have not used Facebook for years, I am sorry to say, partly because I have very boring friends who post endless political rants, and I got bored of wading through those. Back in the old days, I found that trying to find and use those controls was quite difficult. Does a lot of thinking go into the consumer interface to set my controls?

Richard Earley: Yes, a tremendous amount. I agree with you that it has got a lot better since five or 10 years ago, because we have done a lot more research into what people find meaningful. There is a balance to be struck there between giving people too many options, which means they will be overwhelmed and will not use them, and giving people options they can use and easily understand.

I mentioned the ability to see the newsfeed chronologically or to see from your favourites. That is accessible right from the top-left corner of your screen when you open Facebook on a desktop. You have the ability to give direct feedback on any post that you see on Facebook. You can click on the three dots at the top right corner of any post and say that you do not want to see anything from this person for 30 days, which we call snoozing. You can unfriend them. You can give feedback to the algorithm and say that you want to see less like this. In the announcement just last week, where I said we were testing some new surveys with users to understand what they find helpful, we are testing a new way to make it even easier to do that. Now, next to posts, there will just be a small X. If you see a post and you do not want to see more of that type of post, you just click on the X.

Lord Vaizey of Didcot: That is brilliant. Katy, it took me years to learn how to pin a tweet, but I know how to block and mute. What tools is Twitter coming up with to help me prevent all the abusive tweets I get from my fellow committee members?

Katy Minshall: If you will allow me, I will answer your question in two parts: what user choices you have now and where we would like to be in the future. On the first, we really think about how we can give you as much control as possible over the experience you want to have. When you sign up for a Twitter account, at the outset you can choose which accounts you want to follow. You can decide whether your account is private or public. If it is private, only people you have approved as followers can see what you are tweeting.

There is a whole range of options that you can experiment with. You can use our sensitive media settings if you want to automatically hide any photos or videos that might be sensitive or violent. You can turn off video auto-play if you do not want something to automatically play in your timeline. You can control who can reply to tweets you send out. The one I think is most important, which I use a lot, is the ability to mute words and hashtags, or even emojis. You can see all sorts of day-to-day use of this. If you did not want to see spoilers for your favourite TV show, you would mute the "Line of Duty" hashtag. Equally, there are really important uses. If there are certain conversations that you find distressing or that you do not want to see, you can mute the words associated with them, so they do not appear in your home timeline.

There is also a question of what kinds of user choices we want to be able to give people in the future. This is where algorithmic choice is fundamental. Like Richard was saying, on Twitter you can turn the algorithm off. If there are any committee members here who are on the app on your smartphone, you go to the top right-hand corner of the home timeline. There is a sparkle icon and you can turn the algorithm on and off at will. If you have it off, as Richard said with Facebook, you just see tweets from accounts you follow, in reverse chronological order.

I have the algorithm on, because if I am diving into Twitter in the afternoon I want to quickly see the tweets most relevant to me. It would be really interesting for the future if it was not just a binary choice. What if there was a range of algorithms that I could select from?

That sort of question, and the more fundamental questions of transparency and algorithmic explainability, is what our responsible machine learning initiative at Twitter is looking into. They are looking at how we can even think about third parties developing algorithms in years to come and, critically, how we can be transparent throughout. They will publish some analysis that they have done on some of the risks associated with the algorithms we use every day on Twitter, which we hope will improve our collective understanding of how these different processes are working and the effects they are having.

Lord Vaizey of Didcot: You have set a hare running with "Line of Duty". This discussion is too civilised. It needs to get angry so that, at some point, one of my fellow committee members can say, "Joseph, Mary, Jesus and the wee donkey". Thank you for explaining what that star is on Twitter, which I never knew and have just clicked on.

The Chair: Let us move on from design to moderation.

Q209 **Baroness Featherstone:** Hello to both of you. We have heard from witnesses that it is not terribly clear—in fact, it is unclear—who makes the decisions on content moderation. Do those responsible for content make the decisions up to a certain level? If the decisions get harder, is that escalated higher up to the CEO or the business manager? Talking to David Kaye of the UN, he posed the issue: do you work under the UN guiding principles on business and human rights, which are to protect everybody's rights, or are decisions made, occasionally, sometimes or often, on a political or business basis? In short, we want to understand to what extent those who design and implement content moderation policies are insulated, if they are, from commercial and political pressure.

Richard Earley: That is a very big and really important question, so thanks for the chance to speak about it. If it is helpful to you, I will answer the question in two ways. The first is how we design the policies that determine what is and is not allowed on Facebook: the community standards that I mentioned previously. Secondly, I will talk about how we enforce those policies. It is all very well having great policies but, if you cannot enforce them fairly, they are not much use.

First, you mentioned the UN guiding principles on business and human rights. Those are incorporated in, or are part of the inspiration for, our community standards, which, as I said, are the rules we set with external experts to determine what can and cannot be said on Facebook. In February this year, we published our human rights policy, which explains how we use the UN guiding principles and the International Covenant on Civil and Political Rights in the process of developing the policies we have that restrict hate speech, cruel and insensitive comments, promotion of terrorism, et cetera.

In those policies, which are all made publicly available on our website, we follow the precepts from the international covenant and the guiding principles, which are that you should restrict speech only in order to prevent imminent physical harm or damage to someone's reputation, public health or something such as that. Those are embedded in the overarching framework we set for our policies.

To touch briefly on how we write those policies, this is part of the answer to your question of who decides. Facebook ultimately determines what the policies are and we agree them ourselves. When we notice that a change needs to be made in them, which happens very often because the way people use the internet changes all the time, we go through a very thorough process of developing a potential update to the policies. That involves reaching out to stakeholders, people with lived experience and experts in the field all around the world, taking on their suggestions for what changes might be made and synthesising those comments.

Then we have an entire company meeting where we discuss and suggest issues that might arise from changes. When a decision is made, we publish the minutes of those meetings publicly on our website and announce the changes to our standards on the same website. That is the process we go through for determining policies.

When it comes to enforcing them, you asked whether decisions get escalated up the chain when they are difficult to make. That is true in a sense, in that we have more than 35,000 people at Facebook who work on safety and security, including those who are designing the rules and working with experts, as I just said. Around half of them are directly reviewing content. They will review that content based on what type of content it is.

If it is a post that might contain nudity, for example, that can be analysed by somebody regardless of their expertise. When it comes to hate speech or misinformation, we use specially trained teams who are experts in those issues—including, particularly for things such as hate speech, country experts. There is a team of experts on the UK who can give advice to one of the reviewers if there is a decision and they are not sure about the answer.

Increasingly, though, because of the volume of what we see on Facebook, we are using technology more and more to help us make decisions about posts that might break our rules. That is partly because we believe that it should not necessarily be for those who are the targets of hate speech or bullying to have to report what they see in order for us to find and remove it. We have invested a lot of money in technology to help us find and potentially remove, for example, hate speech on our platform. That is really showing results. In our most recent transparency report, which we published in February of this year, we now remove more than 95% of the hate speech on Facebook before anyone reports it to us, using our technology. That is up. It was only 25% about three years ago, so there has been a really rapid increase in us doing that.

As Katy said just now, there is a potential for those algorithms to be susceptible to bias. We have teams in Facebook that look at this and try to analyse the impact of these tools, to make sure they are having a fair impact.

Baroness Featherstone: Are you saying that a decision has never been made for business reasons that, to an extent, contradicted what is within that global control?

Richard Earley: As I was saying before, it is in our business interests for us to have a platform where people feel safe and able to express themselves. That is why we write these policies. It is incumbent on us to keep these policies up to date. The use of technology changes so much that there have been plenty of changes to these policies over time, and there will continue to be. People use Facebook all the time, every day, to connect with their friends and families, to support businesses and charities they like and to campaign on issues they care about.

We have to be extremely careful when making these changes that we are not inadvertently silencing people who are using technologies such as Facebook to campaign for legitimate issues. The "Black Lives Matter" phrase, for example, was first used in a Facebook post. The campaigns and the movement we saw last year could not have taken place in the same way without the power of social media. We take very seriously the need to balance those responsibilities when we are writing rules like that.

Katy Minshall: We make decisions based exclusively on whether a tweet or account breaks our rules. You are right that we ingest those reports in different ways. Like Richard was saying, sometimes it is a user report. If I tweet something to you that is abusive, you might report that and it is reviewed by a content moderator. It might be that an algorithm identifies a tweet that looks like it is breaking the rules. It might be that one of the partners we work with on quite specific issues comes to us with a tweet.

We work with the Department of Health and Social Care and the Government's counter-disinformation unit. They will get in touch, saying, "We think this tweet looks like Covid-19 misinformation". That might mean that specific team looks at that tweet. What is unanimous across the board is that those decisions are made exclusively on whether a tweet or an account breaks the rules.

In how we develop those rules, we try to be as open and consultive as possible. That is easier for Twitter to do, because we are an open, public platform. We have made Twitter data available for a long time now. Tens of thousands of researchers have accessed our APIs over the past decade to conduct research in areas such as online harms. The challenges we see on Twitter are incredibly well documented. That means that it is easier to work externally with safety organisations and experts, who are looking at these decisions as well and can develop solutions with us, whether they are policies, products or anything else.

We have tried to push ourselves over the last few years in asking the public how they would like us to moderate content on certain issues. We have run public consultations on a number of areas of content moderation, such as synthetic and manipulated media, often known as deepfakes. Equally, it could be a photo that has had a slight amendment designed to change the meaning. We ask people how they would like to see Twitter respond to those challenges. We have just closed a public consultation on how we should moderate tweets from world leaders, which I am sure we will get into.

We try to make the rules of the road for developing a really open, transparent and consultive process. When we are making those decisions, based on our rules, they have that credibility and have been designed as effectively as possible.

Baroness Featherstone: Can I be clear, Katy? You are basically saying that at no point will a decision be made based on politics, commerce or business reasons.

Katy Minshall: Yes.

Baroness Featherstone: My follow-on question is to Twitter alone. Excuse me, Richard. Who made the decision to ban President Trump—his tweets, not him? It attracted a lot of criticism. World leaders were worried that it could be held up against the West as us having political censorship, for example. I want to know at what level that decision was taken. To dig down a bit into what you have just been saying, I cannot actually see the difference—I mean, I can see the greater difference—

between removing President Trump and removing Ayatollah Khamenei, who have both said similar and dreadful things much of the time. Why and how was one removed and not the other?

Katy Minshall: This is something we are looking at right now. I can give you a bit of background as to why the tweets you are referring to are on the service and where we are right now. In 2019, we set out the approach we would take to world leaders. We said at the time that some violations of our rules are so severe that we have to take the tweet down, no matter who said it. We acknowledge that there might be times when a world leader tweets something that breaks our rules, but there is a public interest in knowing that that world leader said that thing.

We said in 2019 that at those times we might put a label in front of the tweet, saying, "This broke the rules. We're hiding it from you, so you have to opt in to see it, and we'll restrict it algorithmically, turning off likes and replies". In terms of the tweets you are referring to from Ayatollah Khamenei, we said that, where tweets might be considered sabre rattling, were directed at other Governments and were focused on military or geopolitical issues, we would err on the side of leaving them up. That is for two reasons: first, because of the public interest in knowing that that leader said that thing; secondly, to give people the opportunity to hold that leader to account for saying what they said and being able to respond publicly and directly.

Now, two years on, it is the right time to review whether that is the right approach. Crucially, we have a responsibility that it is not just Twitter setting the rules of the road on such a fundamental issue. That is why we have been running a public consultation, inviting government officials, human rights activists and safety experts around the world to tell us how they think we should moderate content from world leaders.

Baroness Featherstone: To me, looking in, it seemed a great inequality of approach, but I understand what you are saying. Can you confirm that both the comments that were made when President Trump was still on and the decision to ban him just happened? There was no decision by a board, by anyone senior or anything. That just happened because of the algorithms.

Katy Minshall: His account was suspended because it broke our rules. I am sorry; I cannot see a link function here, but I can share with you the analysis that we put on our website as to why we decided that the tweets he shared broke our rules and were subject to suspension. We recognise the really substantial public interest in knowing why we took the action we did, which is why we published it and made it available.

You are absolutely right. These are such fundamental issues and now is the time to look at the best approach to how we moderate tweets from world leaders.

Baroness Featherstone: Thank you very much for that. We will be interested to see where you go on that, because it is, as you say, a fundamental issue.

Q210 **The Chair:** Let us actually see what Ayatollah Khamenei said. He said, "#Israel is a malignant cancerous tumor in the West Asian region that has to be removed and eradicated: it is possible and it will happen". Your tests seem to be a distinction between sabre rattling and incitement. As it stands now, and I understand that you are reviewing your rules, Twitter does not see the statement I have just read to you as incitement.

Katy Minshall: There are two things. First, you are right. That tweet is still on the service. We have made a determination that it is in the public interest for people to know that he tweeted that thing. Secondly, in our world leaders policy, which we shared in 2019, we said at the time that we would not backdate it. I believe the tweet you are referring to is from 2018.

The Chair: This is one from 2020: "We will support and assist any nation or any group anywhere who opposes and fights the Zionist regime, and we do not hesitate to say this". Does that predate it? Is that not incitement?

Katy Minshall: It does not predate it, but that is an occasion when we have assessed a tweet against our approach to world leaders and erred on the side of leaving that tweet up, so that people can hold that leader to account for saying what he said. You are absolutely right: these are really important questions, which is why we are running a consultation on whether the approach we have taken is the right approach.

The Chair: I will read you another tweet: "Palestine will be free, while the fake Zionist regime will perish. There's no doubt about this". In the future, you may take a different view on that tweet.

Katy Minshall: Yes, we may. We do not have a predetermined outcome in mind with the consultation we are running. We are really interested to hear how people think we can most effectively moderate tweets from world leaders.

The Chair: Let us move on to the role of fact checkers. We will stay with Twitter for a moment. We met some fact checkers and they said, "Our role is not to determine the absolute truth in every case". Basically, fact checkers are journalists who are checking the work of other journalists. They say, "Our role is to provide context to stories about controversial issues, so that readers"—your users, for example, if you use the fact-checking service—"can make up their mind about an issue". Is that how you use fact checkers?

Related to fact checking, last month, Twitter, you marked a tweet by a professor of medicine at Harvard Medical School as misleading. You limited the users' ability to share that tweet. The tweet was, as I say, by a professor of medicine at Harvard Medical School. Who in your organisation would have been qualified and who would have decided that a professor of medicine was wrong?

Katy Minshall: It is not Twitter saying he is wrong or misleading. It is the CDC and health authorities around the world. From my understanding, the tweet you are referring to says that, if you have had Covid-19 before, you have natural immunity and do not need the

vaccine. That is different to what the CDC and other health authorities around the world have said, which is that vaccines are effective for most people. When people see that tweet, we want to direct them really quickly to authoritative sources of information, such as the CDC, the NHS or the Department of Health and Social Care, so they can see what the official guidance is and make up their own mind.

The Chair: On these highly controversial, really current issues of public health, you think there is a danger in having debate among acknowledged experts. It is far better if everybody just sees the official public health position, even though that may change in time.

Katy Minshall: It is a good question. That tweet highlights the complexity of trying to moderate the Covid-19 conversation. You are right: on the one hand, the information environment and what is accurate with regards to the pandemic is ever evolving, with Governments providing different and sometimes competing advice. You saw that with the AstraZeneca vaccine.

The Chair: Professor Kulldorff, the person we are talking about, is a member of the US Centers for Disease Control and Prevention and a member of its vaccine safety technical sub-group. He seems to be at the heart of a public health authority to me.

Katy Minshall: There is a real public interest in health officials all over the world engaging in good-faith debate on the efficacy of different interventions. That has to be balanced by the fact that people will come to Twitter looking for the latest official, accurate information about the pandemic and the vaccine. We want to point them really quickly in the direction of those sources as much as we can, so that they have access to all the information and can make up their own mind.

The Chair: I am not saying this pejoratively, but just to be clear about your position. It will stifle debate among experts, who may differ from every single last word of the official guidance.

Katy Minshall: Critically, we did not take that tweet down. It is still on the service. People can see it, tweet about it and offer their own opinion. Health authorities have said it is misleading, so we want to ensure that people are also presented with the option of seeing the official guidance, when it is different to what that tweet says.

Q211 **The Chair:** Facebook, you marked an article by Carl Heneghan, professor of evidence-based medicine at Oxford, as false. Could you describe the process that was gone through in your organisation to come to the view that it was false?

Richard Earley: Similar to what Katy has just been describing, it is not Facebook, the organisation, that determines the falsity or otherwise of posts like that. We have a network of independent, third-party fact checkers who we work with. There are more than 80 of them around the world and we have three here in the UK. When those third-party fact checkers see a post that they believe is false or misleading, or when a post on Facebook or Instagram is getting a lot of reports for being false,

or a lot of comments underneath it that indicate it might be false, they carry out the work that they do to rate the accuracy of the post. If they determine it to be false or misleading, we take a number of actions on it.

One of them is, as you just described, to cover it with a shield, which says that there may be more information regarding that topic. That includes a link to the article from the fact checker, explaining why they have reached this conclusion. Secondly, we show it much lower in people's newsfeeds and elsewhere, so we reduce the distribution of it, normally by up to 80%. Lastly, we give people more information about that content.

The Chair: Could you tell me how qualified an expert the fact checker would have been to analyse all this other information and the view of other experts, in order to come to a view that what Carl Heneghan, professor of evidence-based medicine at Oxford, had said was false?

Richard Earley: All the fact checkers who work with us at Facebook are certified by the International Fact-Checking Network.

The Chair: Will they have had any medical or scientific qualifications, or will they have been a journalist?

Richard Earley: It depends on who did the fact check. As it happens, the point you are raising is a very good one. We know that fact checkers are incredibly underresourced and we rely on them tremendously for the work they do to help keep our platform safe.

The Chair: Fact checkers tell us that their job is not to determine the truth. Their job is to provide context for people to make up their own minds.

Richard Earley: That is right. That is why last year, together with the International Fact-Checking Network, we sponsored a £1 million bursary to put dedicated health experts into fact checking networks, including one here in the UK, so that they can bring their expertise to those positions. Much like Katy said, we are not removing the post from Facebook. We are providing this additional information on it, in order to give people the additional context, as you just described.

The Chair: In this case, you did not give additional information. You described his article as false.

Richard Earley: I cannot see the article you are describing, but it the shield that we put over the article should say, "This article has been rated false. Click here to learn more". Clicking on that button will take you to an article written by the fact checker that explains their reasoning for rating it that way. We will also put around that article, when you see it on Facebook or Instagram, additional information from news organisations and Governments to provide more context to the story.

Q212 **Baroness Buscombe:** I always get quite concerned when one talks about keeping a platform or some aspect of social media safe. If that was the case, why are people across the world clamouring for legislation to regulate what you do? It is a huge ask to suggest that anybody is

keeping anything safe. There has to be an element of risk here, surely. I will press you a little further, particularly you, Richard. It all sounds very slick, to the point that my colleague Lord Vaizey was making earlier. Things are not all right. If they were, we would not be having this conversation. We are looking at how we protect freedom of expression online.

We are talking about truth, to the best of our ability. Mark Zuckerberg said in May 2020 that he did not think that Facebook, or any internet platforms, should be arbiters of truth and that there should be particular tolerance for political speech. Yet, in October 2020, Facebook reduced the number of people who would be shown a story, generally seen as a pretty negative story, about Hunter Biden in the *New York Post*. Facebook's justification was that the story was misinformation, but we have seen no evidence that was produced to that effect. Maybe you could put us right on that. Surely this is a case of Facebook, or one of your 35,000 trained teams and country experts, being the arbiter of truth on this. How would you respond, Richard?

Richard Earley: There is quite a lot in your questions, which I would love to address. To go straight to the question you asked at the end, the action that we took towards the *New York Post* story about Hunter Biden was in keeping with the process that we had set out one year before the US election date, in November 2019, explaining how we would protect the election. When we have signals that a post or story might be false, and it is going viral and getting a huge amount of engagement, we have the option to temporarily reduce the distribution of that story for seven days, to give our fact checkers, whom I was just describing a moment ago, time to investigate and assess the story.

That is because we know that fact checking, as we have said, is a very difficult business. It can take some time to get to the bottom of it. However, slowing that article down does not mean taking it down. As I said before, that article remained on the platform for those seven days, and could be and was still discussed on Facebook, as well as being discussed at length in other media. In the course of that seven days, our fact checkers did not provide a fact check on it, so, after seven days, those demotions were suspended.

While that is an intervention from Facebook, it is an example of a non-binary approach, moving away from the "leave it up or take it down" approach, which has been simplistic in the way it has been applied in the past. I think more of those sorts of approaches will be called for in the future.

On your introduction about clamours for legislation around the world and things not being right, I fully agree with you. I would not want to sit here and pretend that we are saying that there is no bad or harmful content on Facebook. We know, because of the size of the platform that we operate, that there will always be people who seek to use our platform to do harm. We have a responsibility to do everything we can to prevent that.

As you said, part of that is about what we can do to set new rules and enforce them, but we feel uncomfortable about taking so many decisions, particularly on issues such as this about free speech, on our own. That is why we are one of those voices that you mentioned calling for legislation. We welcome the online safety Bill in the UK.

In the absence of that, we are not waiting for legislation. We have tried to bring in additional voices to the debate through the process I described, whereby we bring in experts to change our rules, and the creation of the oversight board. I know you have had evidence from a member of the oversight board before. In the analogous case to the one Katy was describing with President Trump, we have referred that decision to the oversight board to get its views.

It is important to remember that, although there is this view that things are not right and there is a lot of terrible content out there, it is an extremely small proportion of what is on Facebook and being seen by our users. The vast majority of people are using Facebook to campaign on issues they care about, to raise money for charities they support and to connect with their families and local communities. In the pandemic, we saw 3 million people in the UK joining Covid mutual aid groups here, to help support their neighbours and friends who were shielding.

We really want to be transparent about the scale of the problem and do what we can to stop it, but our own analysis has shown that actually only 0.07% of all the views of posts on Facebook contain hate speech. That is around seven or eight in every 10,000 views. The amount of bad posts we are talking about is very small. It is important to keep that in mind.

Baroness Buscombe: I appreciate that. You have a complex array of uses and everything else. I get that. The family and friends thing is a critically important part, but then there is the politics. Can I press you once more? On the one hand, you are reducing what we can learn about Hunter Biden, because you want to fact check that. Meanwhile, Facebook has not, we believe, fact checked or limited the reach by Chinese state media, which denies the human rights abuses against Uighur Muslims in Xinjiang. Why is that okay? This is not very consistent, is it?

Richard Earley: I am grateful to the committee for sharing with me a few of the posts from the page in question, which is the *Global Times*. The situation in Xinjiang is incredibly serious right now. I know that just last week Human Rights Watch published a report describing what is happening there as crimes against humanity. This is something that our company is taking an extremely close look at and is watching very seriously.

We apply the rules that we have been discussing so far to pages and to the posts that they put up. If a page continues to break our rules, either by putting up content that is against our policies or by behaving in a way that is misleading about who they are, or if it represents a terrorist organisation or a designated hate group, we will remove it. In this case, there is a slightly special case for the sorts of organisations that are state-controlled media entities. They combine the influence of a media organisation with the power of a state.

We wanted to go an additional step to make sure that our users know who is behind the stories and the posts they see. That is why last year we started to label pages that we had determined were partially or wholly editorially controlled by a Government as such within our app. You now see, within the "about" section of the app, that that page is state-controlled media from China. Increasingly, we are rolling out the places where this is visible. Now you will start to see these sorts of posts, even underneath the page name itself, whenever it makes a post.

On the Hunter Biden part of the question, I should mention that Facebook was warned in the autumn of 2020 by the FBI in the US that there was a risk of these hack and leak operations, whereby hackers from inside or outside the country would try to disrupt the election. That is why we had that policy in place that I described: to proactively reduce the distribution of posts where there were concerns about their falsity or truth.

The Chair: It is very interesting. I do not want to go into all these individual cases, because I recognise that each of the cases we have discussed has its own background. They are all, in themselves, very serious issues. What we are drawing out here is that we are looking not just for openness in your moderation process, but for a real acknowledgement that consistency is important too. As you develop your moderation processes, particularly bearing in mind the importance of protecting freedom of expression, people using your platforms are looking for you to take a consistent approach across the piece. That is what we are drawing out here.

I am not arguing that President Trump should have been banned or not, or that Ayatollah Khamenei's tweets should have been banned or not. It just strikes us that, as you develop your codes and practice in this area, consistency is critically important. Let us move on to a final question from the Bishop of Worcester.

Q213 **The Lord Bishop of Worcester:** Thank you to both for all the evidence you have given so far. I want to focus upon the area of legal but harmful. We have had a lot of talk within the committee and from witnesses about the proposals in the forthcoming Bill for legal but harmful. Some of our witnesses have been very concerned about what is being proposed, given that the scale of online user-generated content will require platforms to use algorithms for content moderation. As we all know, they are not good at identifying context, nuance or irony.

Some, including Google, have said that if a Government believe that the category of content is sufficiently harmful, they may make the content illegal directly, through transparent, democratic processes, in a clear and proportionate manner. The danger is that, on the one hand, one wants to protect people from harmful content, but, on the other, one wants not to limit freedom of expression. A worry that has been expressed is that, if a penalty is applied to platforms that take down things that it is felt should not be taken down, they will be subjected to penalties as well. You are aware of all this.

I have read your written submissions. Richard, the Facebook submission says a bit more than the Twitter submission, but it seems to be a little agnostic on this. "This is all very difficult. We'll wait and see what's proposed". I would like to tease out your feelings.

Richard Earley: Unfortunately, it is very difficult. Overall, we have been very welcoming of the Government's online harms and online safety proposals, as they have developed. We are looking forward to seeing more detail soon about what will be in the Bill and how the relationship between the Bill, the secondary legislation that establishes the priority categories of harm and the work by Ofcom eventually to enforce the duty of care will fit together.

On the legal but harmful piece, the challenges you have eloquently described frame very aptly the reasons why the Government must consult very carefully on this part of the framework. This is a new approach that has never been tried before. The idea of holding companies to account for policing legal speech raises a lot of questions for us at Facebook. There is a lot of expertise in Ofcom, and we are very pleased to see Ofcom be nominated as the eventual regulator, but there will be a lot of difficult work to understand what the requirements are.

You said in the opening to your question that algorithms are very bad at detecting nuance. That is true, and that is why we do not use algorithms for that kind of thing. At Facebook, we will always use algorithms for what they are good at, which is surfacing potentially bad content and acting on it where it is very clear. Where it is not clear, that is provided to trained reviewers and experts with context, who can then make decisions on whether something is harmful.

We, as social media companies, have to balance those questions of freedom of expression and keeping people safe all the time, with the millions of reports that we receive every day, but it is not a zero-sum game. By preventing people seeing hate speech, attacks or bullying, we give people more freedom of expression and encourage people, particularly from communities that have traditionally been the target of abuse online and elsewhere, to speak up. Getting that right really matters for them too.

The Lord Bishop of Worcester: Can I press you? You say that it needs the human moderation. You mentioned fact checkers. I think you said that there were 80 over the world and four in the UK. When you think of the amount of content that might be judged harmful, although it is legal, would it not require an enormous amount of policing? Is that practicable or feasible?

Richard Earley: We wait to see the details of what the Government have in mind and how it will be interpreted. There are a tremendous amount of posts on Facebook that are not illegal but break our community standards. We already enforce our community standards far above the legal definition of what is hate speech or incitement, for example. We do that through balancing the use of algorithms for what they are good at and humans for what they are good at.

For fact checkers, you are absolutely right that it is a real bottleneck in some senses. We need to rely on these incredibly hardworking people in order to make these decisions and give this extra information. Increasingly, we have been scaling up from the work they do in order to supercharge their work, if you will. Now, when a fact checker rates a piece of content as false or misleading, we can use technology to apply that rating to other posts, pictures or comments that contain the same claims, for example.

In the course of Covid, when there are what we call widely debunked hoaxes, well-known claims that are false about Covid, we have started to roll out the use of technology to automatically apply the sorts of down-ranking and labelling approaches that I described earlier, without one of our fact checkers having to actually assess the content. In the course of the pandemic, while we have removed around 12 million pieces of content for breaching our rules on harmful misinformation, we have reduced the spread of more than 160 million pieces. Obviously there have not been 160 million individual fact checks by our fact checkers, but we have used this technology to scale it up. Technology has a really crucial part to play in enforcing the eventual rules that will be put in place by the online safety Bill.

Katy Minshall: You are right to ask this question. It is a challenging area. It is good that the online harms proposals have designated Ofcom as the regulator. It is a credible, experienced regulator that can bring its knowledge to bear on these really tricky issues.

At the same time, if there is a type of speech that is legal but in which government wishes us to intervene, government should define exactly what that speech is. The risk of not doing that is not just that Twitter and companies like ours struggle to interpret it, but, more importantly, that users of internet services do not know what they should expect from us either.

The Chair: We have run out of time, sadly, on a very interesting and informative question. If either of our witnesses has further thoughts on that, we would be delighted to see further evidence in writing. Katy Minshall and Richard Earley, thank you very much indeed for coming and giving evidence to us today. We will be publishing a report in due course. We are getting towards the end of our evidence sessions and the evidence you have given us today has been very useful to the committee indeed. Thank you for your time this afternoon.