HOUSE OF COMMONS

Science, Innovation and Technology Committee

Oral evidence: Social media, misinformation and harmful algorithms, HC 441

Tuesday 25 February 2025

Ordered by the House of Commons to be published on 25 February 2025.

This evidence session includes content that some may find distressing.

Watch the meeting

Members present: Chi Onwurah (Chair); Emily Darlington; Dr Allison Gardner; Dr Lauren Sullivan.

Culture, Media and Sport Committee members present: Liz Jarvis; Paul Waugh.

Questions 88 - 179

Witnesses

I: Chris Yiu, Director of Public Policy for Northern Europe, Meta; Ali Law, Director of Public Policy and Government Affairs, UK and Ireland, TikTok; Wifredo Fernandez, Senior Director for Government Affairs, X (formerly known as Twitter).

Written evidence from witnesses:

- Meta
- TikTok
- X

Examination of witnesses

Witnesses: Chris Yiu, Ali Law and Wifredo Fernandez.

Q88 **Chair:** Thank you, and welcome back to the second session of the Committee's inquiry into social media, misinformation and harmful algorithms. I start by welcoming Paul Waugh MP and Liz Jarvis MP, who are both members of the Culture, Media and Sports Committee and are taking part in this evidence session as guests.

This is the second session. The Committee invited Elon Musk to give evidence, but he did not respond formally. We are very pleased that we welcome a representative of X this afternoon, as well as representatives of Meta and TikTok. Those three companies, and their subsidiaries, represent a significant proportion of social media traffic and therefore have a significant contribution to make to the debate on misinformation and algorithms.

I note in passing that this is an all-male panel. We try to avoid all-male panels, but given the challenges in bringing the panel together we thought it was desirable to go ahead. We are very pleased to see you and that you were all able to join us. We know that it has been challenging.

The reason why we are so pleased to have you here and have worked so hard for this panel to take place is that there is a strong public interest in misinformation, algorithms and social media. It surprised me just how much public interest there is. Not everything about technology and science cuts through to the public, but there is a strong public interest in this. It is on that subject particularly that I want to thank you for taking part, because of the strong interest that we have. We are seeking to understand what happened with the spread of misinformation over the summer, and to ensure that something like that can never happen again.

I will introduce you as I ask the question of each of you. The interest in this, as I have said, has been significant. I start with Wifredo Fernandez, who is the senior director of government affairs for X. I will go across the Committee to Chris Yiu, who is the director of public policy for North Europe for Meta. Alistair Law is the director of public policy and government affairs for the UK and Ireland for TikTok.

We have a lot of interesting questions and three members of the panel. I will ask you to be brief. I may have to use the Chair's privilege to hurry you along at times. Please don't take that the wrong way. Can I ask each of you whether you agree that big tech companies have a responsibility to be transparent and accountable before Parliament?

Wifredo Fernandez: Thank you, Chair, and members of the Committee. It is an honour to be here. We absolutely agree on transparency. We want to make sure that all our users experience X in a safe manner. We are totally with you in that charge.

Q89 **Chair:** Thank you very much, Mr Fernandez. Mr Yiu?

Chris Yiu: Again, thank you for having us here. We are grateful for the opportunity to come and discuss these important issues with the Committee. I absolutely agree that transparency is incredibly important. It is something that we strive for through all the work that we do.

Ali Law: I echo my co-panellists' thoughts on welcoming us here today. We really appreciate the opportunity. I completely agree with your point on transparency. For us, this is the 12th Select Committee that we have attended in the five years that we have been operating and had a presence in the UK. We are very happy to be here.

Q90 **Chair:** By transparency, my question was specifically about accountability to Parliament. That was the spirit in which you answered it, so thank you very much.

I will move quickly to the actual meat of the subject that we are looking to discuss today. In one sentence again, would you say whether you profited from the riots of the summer of 2024? Would you prefer to be Mr Fernandez, or Wifredo?

Wifredo Fernandez: Whatever the Chair desires.

Chair: We will start with Wifredo.

Wifredo Fernandez: Thank you, Chair. Whenever we have a high-profile incident—a flashpoint, such as that one—it is common for a lot of advertisers to pause and see how the events play out on the platform, and then eventually resume operations. I wouldn't say that we profited from that moment.

Chris Yiu: These are incredibly difficult situations that we are talking about. I think all of us were really shocked and appalled by what we saw. Let's be very clear. The harmful content on our platforms that you are describing is not anything that we want to see. We do not want it. Our advertisers do not want it on the platforms. Users do not want to see it. Categorically, no, it is not something that we want.

Ali Law: There is one thing to make clear. TikTok differs perhaps from some of the services that we are talking about today. We are much more of a global entertainment platform than we are the digital town square. In general, the types of content that we saw around that were probably lower in prevalence and not necessarily particularly misinformation-focused. Having said that, we stood up a command centre of more than 100 people across 10 teams across the world, precisely to take action against any violative content that we found. I echo my fellow panellists: we didn't profit from it, no.

Q91 **Chair:** Thank you very much for those responses. I am going to ask you three questions before we move into more detail. I want to understand, from each member of the panel, what lessons you have learnt from the riots and what you would do differently, if anything, if it happened again. Do you think you dedicated enough resource to coping with the spike in

mis- and disinformation? How did you engage with Government and law enforcement? How could this have been improved?

I will repeat the questions, as there is a lot there. Again, is it okay if I start with Wifredo?

Wifredo Fernandez: Sure. Immediately following the attack, we met with the national security online information teams, the Secretary of State for Technology and Ofcom. We also worked with law enforcement to give them best practices training on our law enforcement portal, which was critical immediately following the attack.

Q92 **Chair:** Did you dedicate enough resources, in your view? Is there anything you would do differently?

Wifredo Fernandez: With every incident, we learn. With every response to such an incident, we learn. We have very clear protocols in place as to how to deal with this content and the challenges that arise when we have the aftermath of an attack like that. Part of that protocol, first and foremost, is understanding how we can prevent harm and what we need to prioritise, understanding the contours of the conversation and where our teams can focus their monitoring and proactive detection. We make sure that detection continues until the harm subsides.

Chris Yiu: Safety is our No. 1 priority. We have very strict rules and policies around what is and is not allowed on our platforms. We have enormous efforts in terms of technology and human efforts to enforce those policies properly. We have 40,000 people at Meta working on trust and safety.

In addition, in instances like those we saw last summer, we have crisis protocols that we put into play immediately. As a result, we actioned a significant volume of content on our platforms. We took down 17 pages that violated our policies on inauthentic behaviour, which is where somebody sets up a page purporting to be someone, perhaps in a local community, who is not representative. We took down 24,000 posts under our policies for breaking our rules on violence and incitement. We took down another 2,700 posts breaking our rules on dangerous organisations. All of that was action that we took proactively, to make sure that our platform was safe and that the issues we are describing were being dealt with. We had direct and, I think, constructive and regular contact with the Government and law enforcement, which is incredibly important in fast-moving situations like those.

On your point, Chair, around lessons to be learnt from these situations, one thing that is challenging is that in fast-moving incidents it can be difficult to establish the facts on the ground in real time. I think that is something which we all need to reflect on and understand how we could do better.

Q93 **Chair:** Thank you very much.

Ali Law: As I mentioned, when high-risk events occur, like unfolding public safety emergencies, as part of our crisis management protocols we launch command centres. We launched one in this instance with more than 100 people across 10 teams, working on a 24/7 follow-the-sun basis to make sure that we were moderating around the clock.

As I said, the vast majority of content that we actually saw during the protest was not violative. It was often documentary or bystander footage, but we saw some rise in some violations, particularly around hate and misinformation. In the two main weeks of the riots, our T&S teams took down tens of thousands of videos that violated our community guidelines, and tens of thousands of violative comments as well.

You asked about our engagement with Government and other authorities. From a law enforcement perspective, we were proactive in contacting and offering assistance with law enforcement. From a community groups perspective, we made sure that we were in contact with groups like CST and Tell MAMA, to make sure that we had further levels of information. Like others on the panel, we spoke to DSIT, the Secretary of State, and at a working level, and also the Home Office on that basis.

To your question on learnings, there are probably three things that I would take away. The first is very similar to the point that Chris just made. Sources of authoritative truth were difficult to ascertain, particularly given the fast-moving events. As we plan our activity and try to establish the veracity of potential claims, that is an area that we would like to both work on and work in partnerships with others on.

When it came to misinformation, we frequently did not see things that were being created on our platform, but did see some level of off-platform co-ordination that then came on to our platform. From our learnings, we need to make sure that we are fully abreast—we have teams that do this—of potential co-ordination on an off-platform basis that may manifest itself on the platform in some way: the idea of sharing video or sharing livestreams.

The final point is a continued commitment to work both from an industry perspective and with the wider value chain when it comes to potential flashpoints. By wider value chain, I include news broadcast media as well. Sometimes we would see a feedback effect towards coverage on there and what we saw on the platform.

Q94 **Chair:** Thank you all for your responses. I want to clarify something, Alistair. You said there was a team of 100 people set up around the world specifically to deal with the misinformation and disinformation that was coming out around the time of the Southport protests.

Ali Law: Specifically to deal with the Southport protests and the subsequent civil unrest, yes.

Q95 **Chair:** Did other members of the panel give a similar resource allocation?

Wifredo Fernandez: I don't have a specific number, but there is a cross-functional team that works around the clock 24/7, similar to this, for a duration or period while the incident protocol is in place.

Chris Yiu: Similar. As I described, we had a crisis response protocol initiated. We had teams working round the clock. I don't have a specific number for you.

Chair: Perhaps it is possible to identify the numbers. We have one example, so it would be very interesting to have something to compare with. Understanding how the crisis response teams work would be helpful to our inquiry. You have all touched on issues that we will want to come back to about the nature of the content that was shared about misinformation. Paul has a specific question.

Q96 **Paul Waugh:** Mr Fernandez, why have far-right influencers like Tommy Robinson and self-styled misogynists like Andrew Tate—both of whom have spread harmful content online—been readmitted to X?

Wifredo Fernandez: Shortly after the acquisition in late 2022, we instituted an amnesty programme that allowed accounts which had previously been suspended to reapply for reinstatement, so long as the violations did not violate the law. Those accounts were included in that programme.

Q97 **Paul Waugh:** Andrew Tate once tweeted, "If you put yourself in a position to be raped, you must bear some responsibility." Would that be acceptable now on X?

Wifredo Fernandez: I can't speak to specific hypotheticals. Our teams would evaluate that under our policies.

Q98 **Paul Waugh:** What is your view? It is not a hypothetical. It is a real example. He has said it.

Wifredo Fernandez: I am not a safety professional. It is not what I do, day to day. I would leave that to our experts on the team.

Q99 **Paul Waugh:** During the riots Tommy Robinson was regularly tweeting that Islam was a mental health issue. He shared videos of the disorder and encouraged people to join in the disorder. Would that be allowed again?

Wifredo Fernandez: I can't speak to that specific post. However, if the Chair would allow, I have brought some examples for members of the Committee of community notes that both Mr Robinson and Mr Tate received in the wake of the Southport attack. We saw community notes in action, providing useful context and helpful context for people on X to address potentially misleading information. I have copies of those, should the Chair oblige us in sharing that with the Committee.

Q100 **Chair:** We can accept that as written evidence. How many examples are there?

Wifredo Fernandez: There are five examples of community notes focused on the Southport incident. I have copies for members of the Committee, if they would so like.

Q101 **Paul Waugh:** Mr Law, shortly after the Southport attack TikTok's Others Searched For bar spread misleading content, such as an incorrect name for the attacker. Why did that happen? What have you done to make sure that does not happen again?

Ali Law: The particular example that you are referring to was, as you say, an incorrect name of the attacker in a suggested search bar. To take a step back, search suggestions are generally auto-generated on the basis of what people are searching for. In that particular example, we were notified about it on 30 July. It was removed from the suggested search that evening, and then it was taken down as a search result altogether the following day.

Would I have liked that to be faster? Yes, absolutely, but I highlight two things. The first is that in the immediate aftermath of the attack itself and then subsequent protests, our main focus was on moderating the videos themselves that contained content. There is a search term. It will obviously return videos. Our main focus was on making sure that the video content itself was not spreading any violative content that would breach any of our guidelines. That is the first point.

The second point is that in response to that, during the period of time, we also added search intervention so that, when people were searching for things relating to Southport and things related to protest, either those search terms were blocked or they came up with a result that then linked to further resources, tool and tips around mis- and disinformation, including with some of our fact-checking partners.

Q102 **Chair:** I want to follow up on that because it is a really important point. The Others Searched For feature brings into your TikTok feed things that you have not searched for. You might not be interested in the supposed name of the Southport killer, but it will suddenly appear there. I also note that the function has been criticised for spreading false and harmful information by organisations such as the Molly Rose Foundation, for example, by recommending suicidal material and false election information. Why do you keep that function?

Ali Law: To be clear, when you talk about the idea of suggested search bringing something to somebody's feed, that is not quite the way it works. The TikTok experience is primarily through the "For You" page, which, if you are not a user, is where you land and you scroll through videos. They are not videos that people that you follow have posted, but videos that our algorithm determines may be of interest to you.

The search feature, when you go to search, will indeed, as you say, come up with Others Searched For, or auto-completed search. As I say, it is primarily automatically generated. In some of the examples that you

were just giving, it might be a thing where we are trying to make sure that we stay ahead of developments from bad actors.

If you take the Molly Rose Foundation example, and suicide and self-harm content, the term "suicide" and related types of search terms are blocked and banned on TikTok. If you try to search for them you will be returned a safety centre result and a link to the Samaritans. Obviously, bad actors seek to undertake differentiating versions of that and use algospeak. A lot of our work is about making sure that we are staying abreast and ahead of those kinds of developments so that we can be responsive to any evolution of that.

Q103 **Chair:** It is a good clarification, but I don't think it quite explains why you suggest queries for users based on key words in their searches. The queries that are suggested will not be what the user was originally searching for.

Ali Law: It will often be commonly related to what they have searched for. Again, I highlight the point that I mentioned earlier. We still absolutely care about a search term in and of itself. We moderate it in line with our wider community guidelines. We also have a very clear focus on making sure that when you search on a term, the videos that are returned are not violative of any of our community guidelines, be that on misinformation or be that on suicide and self-harm.

Q104 **Chair:** Would the wrong name of the Southport killer come up again in a similar situation in Others Searched For?

Ali Law: In the example that you talk about, it came up because people had been searching for it. In line with our broad community guidelines, we then went to assess the veracity of it and acted on it. As I said, I wish we had acted on it faster. I think that is a learning for us. That, in and of itself, is a feature that we need to continue to invest in and continue to moderate. We blocked it from suggested search by the end of the same day on which it was reported, and removed it from search terms entirely the next day.

In part, it goes back to some of the points that we were raising earlier about fast-moving events and authoritative truth. We were very focused on protecting things from a short-form video content perspective, making sure claims did not get into that. Search was slightly lower down the priority list. I accept that we should have been quicker. I am hopeful that we will be in any future approaches.

Q105 **Chair:** Are you looking to make changes to the algorithm that determines what the Others Searched For feature serves in order to achieve that, or are you simply looking at how quickly you respond to basically horrific searches like that?

Ali Law: There are two separate things. The first is that we are constantly looking to make levels of change in order to create the safest possible environment for our user bases. The second is that when you

talk about the algorithm that generates that, as I say, it is based on what people have searched. The area where we are looking to push lessons learnt is on the moderation of those terms and making sure that that is as swift as possible. In and of itself, algorithms that suggest a completed sentence on the basis of what people are searching is not necessarily harmful. It is about the content. That is where we apply our community guidelines and rules. That is what we have to make sure is rapid, responsive and enforcing.

Q106 **Paul Waugh:** Mr Yiu, in 2023, prior to the roll-out of end-to-end encryption on Facebook Messenger, there was a total of 36 million reports of child sex abuse made to the authorities by social media companies. That is 110 million images and videos. Can you tell the Committee, in your estimate, what percentage of those 36 million reports came from Meta platforms?

Chris Yiu: I don't have that figure to hand.

Q107 **Paul Waugh:** I can tell you what the answer is. It is 85%. So you were doing a decent job. Now, because of the end of end-to-end encryption used by Meta, the National Crime Agency suggests that they are expecting a decline of 80% in those reports from Meta this year. Is that acceptable?

Chris Yiu: End-to-end encryption is a fundamental technology designed to keep people safe and protect their privacy. We take issues around safety and security incredibly seriously. There are a number of things that we do in terms of co-operation with law enforcement to ensure that we are able to act when, for example, a report is made on our platforms. We can pass that to the relevant authorities.

In some cases, we are able to look at patterns of behaviour. For example, if an individual is setting up a large number of suspicious accounts, we are able to act on that and refer it. We work extensively with a number of partners around safety on these issues. It is important that we address them clearly and effectively. At the same time, as I said, encryption is a fundamental technology that also contributes to keeping people safe and protecting their cyber-security. These are difficult issues to deal with.

Q108 **Paul Waugh:** Isn't it a fact that 20 years ago someone like Gary Glitter had to go to the other side of the world to prey on children and someone like Jeffrey Epstein had to create his own private paedophile island? Now, all these monsters have to do is go and set up a group on Facebook Messenger. Isn't it true that you have turned Facebook Messenger into Epstein's own paedophile island and a place where you can do what you want without getting caught? Isn't it the case that you are defending that in the name of free speech?

Chris Yiu: There is no place for that content in society. It is clearly abhorrent. It breaks all of the rules and policies for our platforms. On issues like these we need a whole of society response to the challenges. The technology companies, including ours, have a part to play in that. I

think we engage effectively and constructively with law enforcement, but these are also deeper issues in society that everybody needs to be engaged in. We want to do our part in co-operating with others.

Q109 **Paul Waugh:** Another big issue for Facebook is the Facebook community groups that many of our constituents are part of. There are tens of thousands of members. Often, their gatekeepers, or individual admins, are two or three amateurs who are expected to moderate that content, whether it is libellous or harmful or other damaging content. Are you satisfied that your guidelines for those individual admins are robust enough to protect the public?

Chris Yiu: We have incredibly strict rules around what is and is not permitted on our platform. Those rules apply regardless of whether we are talking about Facebook groups or other areas. We run the same policies and enforcement against them. Users in those groups can report content that is concerning to them. We are very clear with people about what is or is not expected on the platform.

Q110 **Chair:** People in the group have to report the content. Nobody outside the group can report content.

Chris Yiu: Some groups are set up publicly. Some groups are set up for their members. That does not prevent the members in those groups from reporting the content.

Q111 **Paul Waugh:** You are then putting a huge onus on two or three amateurs who moderate each of those groups, which have tens of thousands of members. You are putting the onus on those individual admins as gatekeepers to see whether or not that content is or is not libellous or harmful.

Chris Yiu: No, because, as I said, we run our own enforcement on the most severe harms and content that is prohibited on our platforms, and we allow users to report. As appropriate, we investigate those and take action where it is required.

Q112 **Paul Waugh:** Do you have any guidelines for those admins?

Chris Yiu: We have the community standards. I am happy to share all of this material with the Committee after the session. We are very clear about what is and is not allowed on the platform.

Chair: One of the issues that Paul is getting at is that this was a conduit for misinformation and for the organisation of the riots during the summer. The extent to which they are managed and controlled, or not controlled, is of concern. That is something the Committee is very interested in. I think Lauren wants to come in on that particular issue.

Q113 **Dr Sullivan:** On the one about privacy that you were speaking about, Mr Yiu, does freedom of speech mean freedom of consequence? In the privacy thing that you just mentioned, are we protecting people who may have said and done some pretty horrendous things? Facebook and other

platforms have our data. You are directing the algorithms to it. It feels like a David and Goliath battle going on. Can you give us trust and confidence about the consequences of the impact of people saying and describing things in the misinformation, and that they will be held accountable and that you will not protect their anonymity by privacy and encryption?

Chris Yiu: As I described, we have community standards that govern our platforms. We have had those for more than two decades. Under those standards we prohibit, for example, terrorist content and content designed to incite physical harm or violence. There are a number of other restrictions designed to keep people safe. We run technology at scale to find the most harmful content and remove it from the platforms. When it is found, as I said, it is taken down. Where there should be consequences for the individuals involved, those accounts may be suspended.

There is a wider point around responsibility. The platforms have a responsibility to set policies, to set the standards and processes around them and to enforce them correctly and to the best of our ability, which is what we strive to do. The wider question of holding people to account for their behaviour and some of the things we have been talking about today does not belong just to the platforms but to society. It belongs to the courts and law enforcement. It is for all of us to work in partnership to make sure that the internet and our platforms are a place that people can feel confident on.

Q114 **Dr Sullivan:** But you won't be protecting people in that sense.

Chris Yiu: As I said, where content breaks the rules there are consequences.

Q115 **Liz Jarvis:** Could I ask Mr Fernandez about X premium accounts? X offers algorithmic prioritisation through the X premium accounts, but what checks are in place to see if a user is suitable for those?

Wifredo Fernandez: That prioritisation is in replies. You have a boost within the reply product surface, so when you comment on someone else's post, depending on your level of X premium, you have a different prioritisation. There is basic premium, premium and premium plus, premium plus being our highest tier that gives you the greatest boost in replies.

Q116 **Liz Jarvis:** But what checks are in place to see if what they are saying is accurate or true, or are they allowed to post what they like?

Wifredo Fernandez: All users, regardless of their subscription, whether they are paid or not, are subject to the same rules.

Q117 **Liz Jarvis:** But what steps are taken if premium users share harmful or misleading content?

Wifredo Fernandez: For misleading content, we have an intervention called community notes. Every post and every user is subject to and

eligible for community notes, including advertisers. This is a decentralised network of contributors who add helpful context. Only when two contributors have previously disagreed and actually agree that a note is helpful is when that note is placed on a post. X has no power to place or remove a note; it is completely powered by people on X.

Q118 **Liz Jarvis:** Community notes don't stop the premium user from sharing harmful or misleading content, do they?

Wifredo Fernandez: No, but independent academic research has shown that accounts or posts that receive community notes are 80% more likely to be deleted; 50% to 61% are less likely to be shared. We see an incredible impact in reducing virality, and posts that receive community notes are demonetised.

Chair: We may as well come now to the points you are making about community notes versus fact checkers.

Q119 **Emily Darlington:** We will start with some blue-tick accounts during the riots: "Illegal immigrants moved into the Dragonfly Hotel, Peterborough," with a video identifying exactly where they were and where protesters were to go. That was a blue-tick account. "The Embassy Hotel in Gateshead houses hundreds of illegal immigrants in hotel next to nursery." Again, that was with a video encouraging people to go and riot there. Is that within your standards, or did you do anything to remove those during the riots?

Wifredo Fernandez: I can't speak to those specific posts; it is possible they may have received community notes.

Q120 **Emily Darlington:** They didn't.

Wifredo Fernandez: I am happy to have our teams have a second look at those posts. I don't have them in front of me, so I cannot tell you.

Q121 **Emily Darlington:** Let's stick with some of the stuff that blue-tick accounts, with the algorithm, therefore boost. Let's start with one we talked about earlier, Mr Andrew Tate's blue-tick account—wanting to make sure many people see them: "Labour cover up rape because they are rapists." Is that true?

Wifredo Fernandez: I am not in a position to evaluate his statement.

Q122 **Emily Darlington:** Another blue-tick account: "Jess Phillips is a rape genocide apologist." Is that true?

Wifredo Fernandez: Again, I am not in a position to evaluate that statement.

Q123 **Emily Darlington:** Let's see if you have an opinion on any of these. Apologies for some of the language coming out here: "This cunt is openly screwing everyone, and he doesn't care why. We rise up and openly shoot Keir Starmer as he needs to step down, or the people of the UK will force him to. If not, I will", swear emoji, swear emoji, swear emoji. Blue-

tick account. Is that within your rules?

Wifredo Fernandez: I will have our teams review that.

Q124 **Chair:** What would a review mean? It is not community notes. If they review that, will they look at taking it down?

Wifredo Fernandez: It depends. I don't have the posts in front of me and I am not a safety team member who evaluates them on a day-to-day basis. To sit here and say that we find and make the right call every time would be wrong. We make mistakes and we learn from them. We did take action on tens of thousands of posts, from account removal to labelling sensitive media to labelling hateful content.

Emily Darlington: I am happy to pass these to you. They took me about 10 minutes to pull together; literally, that is how quick it was. Let's talk to one that was reported and what action was actually taken, or not.

Chair: Can we talk about the speed of action? I think that is very important.

Q125 **Emily Darlington:** In November, I posted a tweet about my petition to save my local post office. This was my reply from one person: "You are a traitor to the British people and you will swing oh so slowly on a gibbet." I had to look up what a gibbet is. It is not a pleasant thing. It was reported as harmful and violent speech in violation of your rules—"Expressing desire for violence is not allowed." That is your rule. Those are your safety rules. No action was taken. No action has been taken. The post is still up. The same account continues to post: "The Jew is the enemy of our humankind. Bend over, bitch. I will fuck you. Here are your shekels. The goal is to replace Europeans with browns before we organise to remove the filth. Zelensky will get his Mussolini treatment for what he's done to Ukraine. They are the enemy within; they must be destroyed. The Jew has infiltrated white society to teach children to hate themselves. Muslims are disgusting homosexuals. Zelensky is a dead man walking." This is all from the account—luckily, not all towards me but clearly all towards somebody. Is this acceptable under the guise of free speech on X these days?

Wifredo Fernandez: No. Those comments are abhorrent and we will have our teams take a look. Absolutely.

Q126 **Emily Darlington:** Can you assure the Committee that this account will be removed?

Wifredo Fernandez: I can't make any assurances, but I can assure you that our teams will review it under our terms of service. I am sorry that you had to experience that.

Emily Darlington: Okay. Let's move on. The thing is that these are not unique. I am not the only MP who gets these kinds of messages on these accounts, reports them and no action is taken.

Paul Waugh: You are happy to make money out of it, aren't you? Isn't that the point?

Q127 **Emily Darlington:** I want to turn next to Alistair and TikTok. Looking at your own guidelines, I want to understand how the algorithm catches this and how it changes it. They say that synthetic or manipulated material showing realistic scenes must clearly be disclosed, and do not allow synthetic media of any real private figure. You allow some "latitude for public figures", but still do not want people to be misled by political or financial issues. It took me about 10 minutes to find an AI of Keir Starmer standing in front of Downing Street and then jumping into a boat full of migrants: clearly AI; no sign or disclosure of AI. This post has been seen 2.8 million times. It has been liked over 244,000 times. It has been shared over 122,000 times. How do your algorithms and your own rules get applied in this situation, and are they working?

Ali Law: First, I would love to pick up with you afterwards that specific thing. I can take that to my trust and safety team and get an evaluation. It is worth us separating two things. There is the moderation approach that we take and there is our algorithm which helps recommend things to individuals. The moderation approach starts very much with AI, and AI being able to use models to screen content that is uploaded for any of our community guideline violations. AI is particularly good at some things; it is good at being able to detect things like pornographic material, blood and that kind of thing. For others, it may refer to human moderators who have to use their nuance, skills and training to be able to rule on other elements that can include hateful behaviour and misinformation and misleading information as well. That is done on upload.

We also make sure that, as content grows in its level of popularity, it is re-reviewed when it hits certain thresholds. It is not just that once you are through you do not have any further moderation. As it hits higher levels, it does. We can take a look and assess that one very specifically under the community guidelines.

Q128 **Emily Darlington:** On this account, it is not their only video; it is repetitive, featuring Keir Starmer in AI doing various things. You cannot reach 2.8 million views without some level of boost from the algorithm, can you?

Ali Law: Let me explain a little bit how our algorithm works. As I said earlier, our algorithm is based on what we call a content graph rather than a social graph. It is not based on who you follow and what they post; it is based on what you like and your interactions with particular videos. If you can imagine what happens, we show you a video and, if you like it by watching it for a long time, sharing it and commenting on it, we say, "You like that. This cluster of users also likes that and they liked these other four videos, so we'll serve up these four videos and see whether or not you like any more of them," and continue to refine on that basis.

I cannot comment on the video specifically, but I highlight two additional things. There are a couple of different community guidelines where we may look at something like that. One is misinformation. We do not allow misinformation that could result in significant harm to an individual or to society more widely. It is worth highlighting that significant harm is the threshold. I am not a trust and safety professional, but we assess it on the basis of significant harm.

Q129 **Emily Darlington:** You say that you do not want people to be misled by political or financial issues. This clearly falls into that category, right?

Ali Law: What I was coming on to say is that from an AI perspective you are absolutely right; we take AI very seriously.

Q130 **Emily Darlington:** Can an algorithm and AI checking, in terms of moderating content—all three companies are using it, but the question is to you—identify and assess AI-generated material without a human factor?

Ali Law: We are putting a lot of investment into making sure that we improve as rapidly as possible.

Q131 **Emily Darlington:** But can it do it today?

Ali Law: I am not a technical person, so I cannot give a definitive answer. One thing to highlight is that we were the first video-sharing platform to adopt standards set out by the Coalition for Content Provenance and Authenticity. When it comes to AI anything that is generated on TikTok is automatically labelled as AI, but obviously sometimes it will be generated by other services. The C2PA coalition is a group of companies trying to create a level of standards such that metadata can be shared so that, even if something is uploaded as an organic video but is not, we have a way of being able to detect that. That is work in progress and we are moving towards it.

Q132 **Emily Darlington:** If people use the TikTok tools to generate the video, it will automatically do that. That is not necessarily how content is created.

Ali Law: That is why we are taking additional steps to try to account for that.

Q133 **Emily Darlington:** I am conscious that I am taking up a lot of time. I just want to ask a question of Meta. In December, Meta submitted some written evidence to this Committee stating that its third-party fact-checking programme was a "key part" of its approach to combating misinformation. How does that fit with the new programme coming in the US and the heavily shared leaked internal guidance suggesting that it will now permit statements such as: "Trans people aren't real. They are mentally ill"; "Immigrants are grubby, filthy pieces of shit"; "Black people are more violent than whites"; and "Jews are flat-out greedier than Christians", clearly promoting racist misinformation?

Chris Yiu: There are two questions in that. In terms of the fact-checkers, we are not ending fact-checking on our platforms; we are changing the way we go about it. We have had feedback over the years that the approach we are taking could be improved and it was not as scalable as we hoped it could be, so we are exploring, as you noted, shifting to a community notes model. We are starting that work in the US. The intention is to deliver a system that is more scalable and has higher trust from our users in a way that still preserves the intent, which is to find and tackle misinformation on the platform.

On the guideline changes you described, I accept that some of those are difficult for the communities that are affected to hear. We have received feedback over the months and years that in some cases areas of debate were being suppressed too much on our platform, and that some conversations, while challenging, should have a space to be discussed. We retain clear rules and community standards prohibiting content that is designed to incite violence. We maintain enforcement around imminent threats of physical harm and so on. I am happy to share all of the details of those policies with the Committee, but we are focusing our efforts like that.

Q134 **Emily Darlington:** You cannot possibly think that statements like those, where there are scientific facts that prove otherwise, are there for anything other than causing and inciting harm, whether online or in-person harm, can you? Do you think it is a genuine debate about whether or not trans people are real, or whether they are mentally ill? You think that is a debate? Is that what you think?

Chris Yiu: We have had feedback over the years that topics that have become part of mainstream discourse—conversations around some of these issues happen among members of the public and happen in newspapers—were being suppressed on our platforms in a way that was too aggressive.

Q135 **Emily Darlington:** You are saying that the statements I read earlier, which are clearly not proven in scientific fact, are now acceptable on all Meta platforms.

Chris Yiu: Where people make statements that violate our policies they will be actioned. Equally, we have had feedback that there are topics where the view was that there should be more room for debate and conversation.

Emily Darlington: I don't even know what to say to that.

Chair: To try to understand what you are saying, Mr Yiu, it is that in certain areas debate around issues like this has extended to a point where people are questioning whether trans people exist, or some people are certainly saying that all migrants are dirty, and that should be allowed because maybe that may be what people are saying. The Committee has questions on that. Do you recognise also that this is not a

debate that is happening; this is you promoting, pushing and amplifying those statements? That makes it a platform for them. It is not the same as a conversation in somebody's home; this is something people will see in their feeds. It was very upsetting to see that during the Southport riots. For many migrant communities in Newcastle and, I am sure, elsewhere, to see all those comments was upsetting and destructive. I am sure you must recognise that.

Emily Darlington: Without being an expert lawyer—I imagine we have some—I do not think these statements in themselves comply with UK law.

Chair: I am not sure about that.

Emily Darlington: I am pretty sure that making statements of antisemitism—

Q136 **Paul Waugh:** Is it a matter for debate that Jews are greedier than Christians? Are you saying that? Is it a matter of debate that black people are more violent than whites? That is a matter of debate according to Meta. Is that what you are saying?

Chris Yiu: What I am saying is that we have had community standards on our platforms for more than two decades. We have revised them numerous times through the years reflecting changes in society and what is discussed.

To your point on the law, we have set community standards and policies that we enforce, and we are clearly subject to the laws of the countries in which we operate. In the UK, that is the Online Safety Act, which is not fully in force but will be shortly, and we expect to be compliant with that. In terms of how we set this up, standards continuously evolve in line with what is happening in society. We want to have a space where people can express themselves and we want to make fewer mistakes around the way we moderate.

Q137 **Chair:** We need to move on. Effectively, you are confirming that leaked description of what was acceptable on Facebook, which the Committee will note. We would also like to write to Ofcom and see whether those statements would be in compliance with future guidelines. If it was illegal content, Ofcom would expect you to take it down. If they are not illegal content, there is still the issue of them being disseminated industrially at targeted communities. I think the Committee as a whole feels quite strongly about that. I want to move on because we have a number of other issues to get to.

Before we come to Lauren, I want to check about Wifredo X's algorithm. We have talked about the blue-tick algorithm. The Committee had a presentation; X made parts of its algorithm public in 2023, I think, and we had a presentation on that. It was quite staggering to see the size of the heterogeneous information network of Twitter, as it then was, which contained 1 billion nodes and described all the links between users,

tweets or posts and adverts. There are up to 100 billion connections within that. That algorithm is effectively run 5 billion times per day. That says to me that X, and I am sure others, are creating a digital twin of each individual on your network in order to target advertisements and content at them for engagement. If that is the case, you must also have responsibility for the targeting of misinformation through those networks. Would you agree?

Wifredo Fernandez: It is a fascinating technical challenge, as I am sure you appreciate. Every day there are 500 million or so posts. Of those 500 million, we are trying to figure out in that moment what is most relevant to you. We start with what we call candidate sourcing. Here are 1,500 posts based on your interests and on communities you may follow and engage with that may be relevant to you. From there, we try to rank those posts in terms of relevance, and then we apply different types of filters. Those can be sensitive content filters, if you have opted not to see sensitive content, and filters to filter out violations of our service. We have posts that are actioned, for example under hateful conduct, that we do not recommend, do not include in search and do not allow to be re-shared, replied to or quote-posted. We filter for those things.

Finally, if you have advertisements in your timeline—if you are a premium plus subscriber you have no advertisements in your timeline—we look at relevant ads that could be served to you, again based on advertisers that are targeting a specific type of content. They have their own controls; they have adjacency controls.

Q138 **Chair:** I think I understand that. I appreciate you going through the process, because it is useful to understand, but the fact is that you have collected an enormous amount of data on each individual—their links and what they liked—which you use to target information, content and perhaps advertising, so you must have a responsibility for the misinformation that comes through to them as well.

Wifredo Fernandez: Yes, and that was why we created community notes.

Chair: I am glad that you accept that responsibility. I think we have discussed the extent to which community notes are effective, and I am sure we will continue to do that.

Q139 **Dr Sullivan:** As was stated earlier today—it is shared across parties—the first duty of a Government is to keep their people safe and protect them from harm. Are social media a threat to the people of the UK, and can you be trusted to mark your own homework and moderate to prevent harm?

Wifredo Fernandez: For us, freedom of speech and expression is fundamental to a democracy, so our goal is to protect that freedom of expression within the boundaries of the law in the jurisdictions in which we operate.

Q140 **Dr Sullivan:** Do you recognise that a potential for harm may come with that?

Wifredo Fernandez: Of course. We understand that every day there are many categories of harm that we have to account for, and that is why we have policies and procedures in place to deal with those harms, based on the risks they pose to our users and their safety every day.

Q141 **Dr Sullivan:** I will follow that up afterwards.

Chris Yiu: Keeping people safe on our platforms is our top priority. I have described some of the rules and policies we have in place and some of the enforcement and technology that we use to do that. We publish extensive transparency reports, documenting our policies and how we enforce them. To give you just a couple of examples, for terrorist content on our platform, which is prohibited, our systems now find and remove more than 99% of that content before it is reported to us, or before someone sees it. To give you a sense of what that means in someone's practical experience, you can find this in our transparency reports where we talk about something called prevalence. Of all the content that you might see on Facebook or Instagram how many of those posts contain content that violates our policies? In this instance it is 0.05%, which is five in every 10,000 posts. That is five more than I would like, but it is a significant effort. You talked about us marking our own homework. We have the transparency reports independently audited by eWISE, so we are not marking our own homework. I encourage the Committee to study those, and we would be happy to follow up if you have other questions on them.

Beyond that, as you will all be very well aware, we and others engaged extensively with Parliament and Government on the Online Safety Act, as it now is. That was truly a momentous effort over a number of years, and has got us to a place where there is recognition that what matters is having the right systems and processes in place and having an independent regulator hold the companies to account. All of that is not yet fully in force, but it will be soon. We have a regular dialogue with the regulator and we expect to comply with it, and I think that is appropriate.

Q142 **Dr Sullivan:** Thank you. Mr Law?

Ali Law: Safety has been our North star since day one. We are a much younger platform than some of our competitors. We only launched in the UK at the end of 2018. From the outset, we included safety by design in creating our product, partly as a competitive advantage. We were coming into a very competitive market and needed to win advertisers and create a good user experience.

We are guided by community principles that feed through into our community guidelines. It is worth just pulling out two. One of those is to prevent harm; that is why we have tens of thousands of trust and safety professionals around the world and are spending $2 billion on safety this year, right the way up to our global CEO, who has safety objectives as

part of his key objectives that he sets for the year. We also have a principle of making sure that we enable free expression, because creativity on our platform is unlocked by free expression. We draw up our community guidelines in line with UN principles on human rights and the ability to express freely.

Like other panellists, we also produce significant numbers of transparency reports and information at our transparency centre. The total amount of violative content—content that violates our community guidelines—that is removed proactively by us is 98.2%. Of that, 85% received zero views and 98% received fewer than 1,000 views. That is not 100% and it is a race that is never run. We will continue to invest in order to try to narrow that even further. When colleagues were sitting here two or three years ago that number would have been in the 80s, so the investment is paying off.

One additional thing to highlight is that we are already regulated by Ofcom. We have been regulated by Ofcom for four years as a video-sharing platform under the transposition of the AVMS directive into UK law at the end of 2020, so we have a very close relationship with them. We meet them very regularly and they have a very firm understanding of our policies, enforcement and approach to wider issues. With that framework, the coming OSA framework and our transparency centre, it is not a case of us marking our own homework. We are driving this activity and enforcement because it makes sense for our users and our platform.

Q143 **Dr Sullivan:** If we can accept that harm occurs on social media platforms, whether that be illegal violent content or algorithms promoting violence, self-harm, bullying online and mis- and disinformation, including fraud, Mr Yiu—referring to Facebook Marketplace and the amount of fraud that has occurred there and has been reported—should that be reported? Do our banks pay for the fraud and it is nothing to do with you, even though you are hosting the potential fraud? Does that make sense? Do you think that is fair? Should Facebook pay an equal share? In terms of the wider harm, do you think social media platforms have a role and a place in restoring and repairing the harm that has been created?

Chris Yiu: Let me take a couple of the points there. Fraud is a concern for everybody. Clearly, we do not want to see it on our platform. It violates a number of our policies. We look for it and we try to remove it, and where it is reported to us we investigate and take the appropriate action. We have a dialogue with Government and with a number of other partners about how we can make progress on that. One of the things that we have done recently is more information-sharing between our platform and the banks, because often it is a complicated chain of events that ends up with someone being defrauded, so it is important that there is more information flowing between the different organisations so that we can learn the lessons and try to stamp it out.

On your wider point about how we engage, we have ongoing dialogue with a number of organisations, as you would expect, across civil society,

academia and elsewhere. We try to contribute to the conversation. We try to explain what is happening on our platforms and we try to learn from and listen to other people. It is right and proper for companies like ours, and the role that we play in society, to be part of that debate. We want our platform to be better. We want our users to have a great experience on it. We want our advertisers to be confident in our platforms and our products. It is squarely in everybody's interests to make sure that happens.

Q144 **Dr Sullivan:** I suppose the real nub of it is that social media platforms make a huge amount of profit off providing content and all the things that we want to see, but there is a darker side with unintended consequences, misinformation and all those sorts of things. Do you think that there is a role, though, for mitigating and using some of that money to help support, protect and repair some of the damage that has been done?

Chris Yiu: For us, the first line of defence is to use the technology to reduce the harm at source.

Q145 **Dr Sullivan:** I am all up for that. Until we are there, and until it is a perfect system where nothing can get through that is not misinformation, what about the consequences to the victims? As you said, you are working with civil society. Does that mean that you are funding some of civil society and you are helping towards youth projects? What does that look like?

Chris Yiu: We have a range of dialogues and partnerships. Over the years, we have worked with a number of different organisations on some of the topics that you have described there and others.

Q146 **Chair:** Do you have a foundation?

Chris Yiu: Sorry, excuse me?

Q147 **Chair:** Do you have a charitable foundation?

Chris Yiu: For Meta specifically, no, we don't, but we have over the years worked with partners and supported different projects.

Q148 **Dr Sullivan:** With money?

Chris Yiu: With money. Those change over time. The important thing is that there is dialogue and relationship between the companies, civil society, our users, our advertisers and businesses more broadly. We are all in this together. It is for us to play our part.

Dr Sullivan: It doesn't feel like that, though, does it, sometimes?

Chair: We are all on this Earth together, but some have far greater power, influence and financial resources than others. That is something that we want to explore in a bit more detail now, Allison, in terms of the business models.

Q149 **Dr Gardner:** If I have time to come back on the generative AI content, I

would appreciate it, but I can jump to the business models. To pick up on a couple of previous questions, can I ask about the TikTok algorithm? You talked about how the algorithm works by looking at likes, shares and time viewing. Human behaviour is such that we would be likely to view sensationalist content for slightly longer. We do not know what your TikTok algorithm is and how those different features and parameters are weighted. You might weight more for the time viewing, and that would be related to advertising revenue, because the longer somebody views content the more money that is made. Are you willing to make your algorithm a bit more transparent so that we can view that?

Ali Law: Let me do a little bit on business model and then algorithm transparency. Just on business model, TikTok is perhaps different from some of our competitors. It is not a case of an advert that is a pre-roll attached to a particular video. Actually, if you are on the For You feed and you are scrolling through videos, an advert will be served as video content in between those, and you can either watch it or you can flick on. That means that there is not a direct link between a huge number of views to a particular video and being able to monetise against that video. Actually, the goal of our algorithm is to provide a positive experience for users overall.

You will commonly find if you are on TikTok that you may flick past a post with thousands of views and then flick on to a post that has very few views, a single-digit number of likes and that kind of thing, because the algorithm is designed to vary your experience, to provide different things that might surprise you or make you curious, where you can discover. Our model is not on the basis of, "Here's a piece of viral content. Let's monetise against that." Our model is on the basis of creating an overall environment that people want to be in, that they feel safe and can express themselves in, and therefore continue to return. It is advertising based a lot more on reach than it is on specific views, if that makes sense.

Q150 **Dr Gardner:** This relates to the infinite content feed that you do.

Ali Law: The scrolling.

Q151 **Dr Gardner:** There is some evidence that there might be harm from that if it creates that dopamine feedback loop. Are you worried about that?

Ali Law: We take the wellbeing of our users incredibly seriously. We don't want to be in a situation where people are using our app and then finding themselves burnt out, tired, or exhausted. It is one of the reasons we worked with the Digital Wellness Lab at Boston Children's Hospital to implement a one-hour screen time limit for under-18 users. It is one of the reasons why we have 9 pm and 10 pm cutoff for our under-16 and under-18 users for notifications coming from the app. We recognise that sometimes people need a prompt to do something else. Our model is based on the idea that people will have a healthy, positive, valuable relationship with the app that means that they continue to come back.

I want to quickly touch on transparency because you mentioned the algorithm and transparency. There are three things. The first is that we have a research API, which more than 600 researchers around the world have access to, that provides data for them to be able to investigate independently trends that they see, content that is becoming popular and that sort of thing. Secondly, we have a number of transparency and accountability centres across the world, including one in Dublin, which the Committee has a standing invite to come and see. Part of that involves having a conversation with the algorithm team about the way that it works. We even have a code room there where you can have a level of review. Finally, we have a relationship in the US with Oracle, which reviews the source code of the app prior to its upload to the app store. We have a number of different touchpoints on which there is a level of transparency.

Q152 **Dr Gardner:** Chris, I am going to come to you. Just before I do, I am hosting a Meta event on Monday via the AI APPG. For transparency, I thought I would say that. This could be one that you all answer, really, but I will direct it to Chris. Does online advertising play a role in incentivising harmful content?

Chris Yiu: No. If anything, it is the reverse. Harmful content is not something that we want to see on our platforms. It is not something that our users want to see, and it is not something that our advertisers want to see.

Q153 **Dr Gardner:** Okay. Simple answer to that one. The United Nations Global Principles for Information Integrity state: "The technology sector has designed digital advertising processes to be complex and opaque with minimal human oversight." I have looked into this with various ad libraries. People can bid for what sites to go to and have their ads on. Viral ads and sensationalist content will be there. Do you agree that it is complex and opaque?

Chris Yiu: Clearly, when the technology is evolving quickly, some of these issues are difficult for non-specialists to get their heads fully around. That is a consequence of how sophisticated the world has become. Nevertheless, we strive to provide as much transparency and insight as we can into the way that our systems are set up. Certainly, for users on our platforms, we provide information in our guidelines. We talk about the different ways that our systems are built and how they work together. We publish things like model cards and systems cards that try in reasonably plain English to describe how different components of the technology work and how they connect together. We try to provide a lot of controls for users on the platform so that they have the ability to tell us in more detail if they want to see less of certain types of content and more of other types of content. For transparency, on Facebook, you can push a button to say, "Why am I seeing this?" You can see some insights into what is causing a particular piece of content to be ranked there for you. We strive really hard to provide transparency for users around this.

Q154 **Dr Gardner:** Can I add one more question very quickly? If there is time to come back to generative AI, I would appreciate it. Wifredo, earlier you mentioned demonetisation. That is one of the recommended standards by the EU—that you demonetise it. During the riots, some far-right accounts were estimated to be making £27,000 from all the generated content. Have you demonetised those? How do you demonetise?

Wifredo Fernandez: Any post that receives a community note is demonetised.

Q155 **Dr Gardner:** Can that process be gamed, where people can make a community note on genuine content to try to reduce the level of exposure to true content?

Wifredo Fernandez: The community note doesn't have any impact on reach. It simply has an impact on the ability to monetise the post. The only way a community note ends up on a post is when two contributors agree who have historically disagreed. It bridges diverse perspectives. It is called a bridge-ranking algorithm and it is completely transparent. All the data is completely accessible. Any researcher can audit the entire history of community notes. We have researchers who are downloading that data every day. We have nearly 1 million contributors around the world and over 70,000 community notes contributors here in the UK.

Q156 **Chair:** I think we established earlier that there weren't community notes on many of the horrendous posts that Emily mentioned, and there weren't community notes on the posts that identified the wrong name for the killer in the Southport attack. If that is your only way of taking money away from them, it doesn't seem very effective.

Wifredo Fernandez: We recognise that there is a challenge of speed and scale, and that is getting better as the system improves and as the number of contributors grows. I have come with specific examples in relation to the incident where community notes ended up on posts. I am happy to share those with the Committee and share others as well.

Q157 **Chair:** Your opening statements were that you did not make money from the Southport riots. Meta has a stock market capitalisation of $1.5 trillion, has total revenues of $164 billion, and 98% of its revenues come from advertising. X has a stock market capitalisation of $9.4 billion. It has total revenues of $2.9 billion, and 90% of your revenues come from advertising. TikTok is privately owned, so it does not have a stock market capitalisation, but 80% of your revenues come from advertising, and you make profits of $15 billion on revenues of $18.5 billion. You are making significant revenues on advertising. Just for comparison, the multiplier between stock market capitalisation and revenue for Meta is 9 and for X it is 3, whereas for Walmart, a retail company, the multiplier is barely 1. You are clearly making money from advertising, and it clearly influences your stock price.

Your stock price is also driven by active users on a monthly or weekly basis, and, as Allison said, far-right influencers, such as Andrew Tate and

Laurence Fox, got 40 million ad impressions on X in the week following the Southport attacks. I think you will acknowledge that there was an increase in activity on Facebook and TikTok. Surely, if your business model is based on advertising and it is driven by activity, you must have made money from there being greater activity following the Southport attacks.

Chris Yiu: We are an ads business, yes. Regarding the content on the Southport attacks and the subsequent unrest that we described, I talked about the efforts that we went to to remove that content from our platform, because we do not want to see that on our platform.

Q158 **Chair:** You may not have wanted to, but you must have done so is my argument.

Chris Yiu: We removed tens of thousands of posts, the vast majority of those proactively. It is not in our interests for our users to have a bad experience on the platform. The last thing that we want is for our users to have a bad experience.

Q159 **Chair:** Wifredo, did X make money out of the increased activity?

Wifredo Fernandez: As I mentioned, it is common for advertisers to pause during heated moments on the platform. We work every single day to earn the trust of our users and our advertisers, and that work continues. When it comes to whether we profited on a specific incident, zooming out, advertisers want their content and their advertisements to be brand-safe, and that is why we put so much effort into controls for advertisers to make sure that their advertisements are around content that they deem safe. That is our goal with advertisers.

Q160 **Chair:** Alistair, quickly.

Ali Law: We are a different kind of service. We are an entertainment service first. It is not chronological posts.

Q161 **Chair:** Eighty per cent. of your revenue is from advertising.

Ali Law: Absolutely from advertising. To your point on a particular high-profile event driving greater activity, we see uptick in activity when the Euros are on or when Taylor Swift was doing the Eras tour. The majority of content on our platform is entertainment, joyful-type content. We absolutely have an incentive to be as robust as we are on enforcing our community guidelines, similar to what other panellists have said. Large brands want that level of safety.

The only other thing I want to highlight is that we operate a different monetisation model for creators. Because we are not doing pre-roll against a specific video, you do not get money as a creator from a particular advert. We have a creator rewards programme, a creativity programme that is designed to encourage creators. You have to be over the age of 18. You have to have a certain level of quality. You have to abide by our community guidelines. The incentives on our creators as

well are that monetisation comes from high-quality content, not just the idea of reach.

Chair: I accept your answers, but given the figures, given the level of revenues and stock market capitalisation and profits, I do not see how you do not have an incentive to have greater advertising spend however it is driven.

Q162 **Emily Darlington:** Two quick points of clarification for Mr Fernandez. First, you said that during points of high tension advertisers tend to pause. Did that happen post Southport? Did advertisers pause on X?

Wifredo Fernandez: It has been widely reported that there was a pause on advertising.

Q163 **Chair:** Could you write to us? You said that you have an API that gives you all that information, so could you write to us with the figures?

Wifredo Fernandez: Okay.

Q164 **Emily Darlington:** I want a point of clarification on your community note system. It says that it involves two people who disagree agreeing to put a community note on. Given some of the posts that I read out earlier, particularly the ones that are inciting violence against women, calling for people to be gang-raped and others, is that considered an opinion where two people do not have to agree? It seems to me, if I understand your community notes properly, that you are saying that Andrew Tate in his calling Labour all a bunch of rapists would have to agree for a community note to go on to that post no matter how many people actually flagged it as misinformation. Is that right?

Wifredo Fernandez: Taking a step back to the system, we have a network of contributors. Those contributors rate notes and earn the ability eventually to write notes, which is a privilege that they may lose. They participate in the system pseudonymously. They start by rating the helpfulness of notes based on a series of heuristics such as whether it is neutral language, whether it contains a high-quality citation or whether it directly addressed the post's claim. Based on the history of their other ratings, that is how you build a profile on that particular contributor. When those two individuals agree that the proposed note is helpful, that is when that note is placed on the post. If another note comes along and it is rated more helpful, it will replace that post. If that note is up for a period of two weeks, it is locked on to that post and cannot be removed.

Q165 **Emily Darlington:** The original poster does not have to agree to the community note.

Wifredo Fernandez: That is correct.

Q166 **Emily Darlington:** Right. How does one become a community note contributor? Are you paid in the role? Is it a human fact-checking role?

Wifredo Fernandez: It is unpaid. Anyone who has an account in good standing with no violations and a verified phone number can sign up to be a contributor. We onboard contributors every week in a fair and randomised process. Globally, we are now near 1 million contributors, with over 70,000 in the UK, and growing. The UK happens to be one of our top five markets where engagement on community notes is high.

Q167 **Emily Darlington:** A final question just to understand how the community note process works. Is the list of your community note people who you have selected to become community—

Wifredo Fernandez: We don't select, to be clear. There is no human at X. It is based on a random number generator that automatically onboards so long as they meet the criteria.

Q168 **Emily Darlington:** Right. You have an algorithm that onboards people. Is that algorithm that chooses and selects available for the public to see, for academic researchers to test and analyse, and can you share it with the Committee?

Wifredo Fernandez: Absolutely. It is publicly available. It is open source technology available to anyone. We see a really great system of software developers who are constantly working on it to help improve it. All the data associated with community notes is publicly available. We update it every single day. We have researchers who are downloading it every single day.

Q169 **Chair:** Can you tell us the proportion who are male or female?

Wifredo Fernandez: We do not collect demographic data on contributors. All contributors are in the public data.

Chair: Okay, thanks very much. We need to move on. We have seven minutes left, and I know that Allison wants to come back. Liz, you have a question as well.

Q170 **Liz Jarvis:** I do. This is on the Online Safety Act, which Mr Yiu mentioned earlier. Can each of you say how the Online Safety Act would have changed events leading up to the riots if it had been in effect?

Chris Yiu: The Online Safety Act is designed around the notion of having good systems and processes in place and ensuring that companies follow those processes and enforce them correctly. That is what we strived to do throughout the incidents last summer.

Q171 **Liz Jarvis:** Okay. It wouldn't have made any difference, is what you are saying, if it had been in effect.

Chris Yiu: As I say, we have the systems and processes in place. Like we were talking about earlier, there are always lessons that one can learn from these experiences. Nevertheless, the systems that we designed and the crisis protocols that existed were put into effect. We had good

engagement with Government and others. As the Online Safety Act comes into force, we expect to comply with it.

Q172 **Liz Jarvis:** Mr Law?

Ali Law: As I said earlier, we are regulated already by Ofcom under the video-sharing platform regime. I echo some of what Chris just said. Our crisis protocols exist to protect users. Ofcom has talked about the potential for looking further at crisis management protocols. We are confident in the steps that we took. We have taken Ofcom through the steps that we took around Southport. They are fully appraised of the mechanisms that we had and the approaches that we used.

Q173 **Liz Jarvis:** Mr Fernandez?

Wifredo Fernandez: We are pleased to have a productive relationship with Ofcom and met with them in the wake of the incident. The Online Safety Act's risk-based approach is consistent with how we deal with these incidents, the protocols that we have in place, how we work to prevent harm and how we prioritise in those moments. Section 22 of the Online Safety Act states: "When deciding on, and implementing, safety measures and policies," there is "a duty to have particular regard to the importance of protecting users' right to freedom of expression within the law." That is a really great framing for how we think about the tension with freedom of expression in moments of heated debate, violence and hatred.

Q174 **Liz Jarvis:** When someone was incorrectly named as the perpetrator of the situation in Southport, it took quite a while to have that post taken down, didn't it?

Wifredo Fernandez: We saw community notes that corrected that. Yes, we recognise that that was an early challenge in the wake of the incident.

Q175 **Chair:** You talked about the tension with freedom of expression, which is something that the owner of X, Elon Musk, has spoken very powerfully about. Do you think that X is a safer online environment since Mr Musk bought it?

Wifredo Fernandez: I do. You can look to our transparency report from last year and the one that is soon forthcoming that will give you a picture of the last year and the actions that we took and the trends in reporting and trends in activities and content violations. You will see that we have a very dedicated and diligent safety team, which we have renamed "Safety" from "Trust and Safety team", because we know that is something that we have to earn every single day from our users and our customers. I am very proud to represent the work of our safety team.

Q176 **Paul Waugh:** How can your safety have increased when the number of antisemitic tweets has risen from two per Jew per year, which is 500,000 per year, to four per Jew per year, which is 2 million a day in the UK? How is that making it safer for Jews?

Wifredo Fernandez: We work on that issue very closely with partners around the globe to understand how they are experiencing antisemitism. It has a few different manifestations when it comes to our policies. It is not just hateful conduct. It is violent event denial. It is abuse and harassment. It is glorification of violence and previous atrocities. We take that very seriously. That is why we continue that work diligently with those partners to better understand it. Clearly, in the wake of 7 October, it is a different reality for those in the Jewish community around the world, both online and offline, and that is a real challenge that we deal with every day.

Q177 **Emily Darlington:** Do you think sharing videos of the Nazi salute on the platform is appropriate and makes it a safer platform?

Wifredo Fernandez: There is a lot of discussion around that. That is why the platform allows for people to debate it and discuss it.

Q178 **Dr Gardner:** I want to jump back to generative AI and the use of AI in creating false images, particularly the ones that are very hard to detect by automated systems such as text data for moderation purposes, and linked to the online harms. People should be able to understand that what they are viewing is fake and is not real; even if it is just filters that make somebody look better than they are, it can affect mental health. Fake imagery of false crimes such as Muslim men chasing a blond toddler is one of the posts that you left up because of external moderation. I know that both TikTok and Meta are looking to do info labels to label the content that is generated. Is that metadata visible to users so that they can assess for themselves that what they are viewing is fake and that anybody sharing clearly fake imagery that is harmful can be held to account?

Chris Yiu: It is important to note that our policies around the content that is permitted on the platform apply regardless of whether the content is AI-generated or not. As you mentioned, where content is generated by our AI tools, we label that accordingly—

Q179 **Dr Gardner:** Is it visible to the user?

Chris Yiu: With a visible watermark on the image and an additional label on the post on the feed. Where AI content is created by tools that are not ours, we will label those as well whenever we can detect the metadata or the invisible watermarks in the content. We label that for users accordingly because we think transparency around this is important. We are working very hard with partners to make sure that can be even more effective, and we are working all the time on technologies to make sure that we are able to moderate our platform effectively. It is a fast-moving environment. I do not think that what we have is perfect, but, yes, we want to make sure that transparency is there.

Chair: We are out of time. I really appreciate your spending time with us. It has been a lively discussion and you have given us a lot to think about. You have given us a lot to take away. I certainly understand better your

relevant processes and algorithms. I am not sure I understand better the future of the spread of misinformation, but that is something that the Committee will consider. I thank you very much for spending your time here. As I said earlier today, I know that one of the things that makes my constituents frustrated is the sense that platforms such as X, Meta, Instagram and TikTok are not accountable and are not something that they have any agency over, and yet have a big impact on their lives. By coming to speak to us today, you have at least supported, as you set out on our first question, accountability towards the British people through Parliament. We really appreciate that, and we appreciate very much the time that you have spent with us.