HOUSES OF PARLIAMENT

Joint Committee on the Draft Online Safety Bill

Corrected oral evidence: Consideration of government's draft Online Safety Bill

Thursday 14 October 2021

9.55 am

Watch the meeting : https://parliamentlive.tv/event/index/f8eb5e28-138c-4ff9-b494-fef97618fb2e

Members present: Damian Collins MP (The Chair); Debbie Abrahams MP; Lord Black of Brentwood; Lord Clement-Jones; Lord Gilbert of Panteg; Darren Jones MP; Baroness Kidron; Lord Knight of Weymouth; John Nicolson MP; Dean Russell MP; Lord Stevenson of Balmacara; Suzanne Webb MP.

Evidence Session No. 4 Heard in Public Questions 92 - 109

Witnesses

I: Guillaume Chaslot, Founder, Algo Transparency; Renée DiResta, Research Manager, Stanford Internet Observatory; Laura Edelson, Researcher, New York University.

Examination of Witnesses

Guillaume Chaslot, Renée DiResta and Laura Edelson.

Q92 **The Chair:** Good morning and welcome to this further evidence session of the Joint Committee on the Draft Online Safety Bill. We are very pleased to welcome our witnesses today: Laura Edelson, Guillaume Chaslot and Renée DiResta. We appreciate that we have witnesses in three different time zones and that you have all made a huge effort to join us at this time, a very inconvenient time to you, so we are very grateful for that.

I would like to start with a question to all three of the witnesses. We often talk about media literacy being an important part of online safety, but for that to work, users need to understand what they are keeping themselves safe from and how they do it. It is hard to do that when the thing that is directing harmful content at you is something that you cannot see and that you have no control over, which is the data that drives the recommendation tools and the online advertising.

In the UK, we want to create an online safety regulator to hold the companies to account for the way they make decisions about content moderation and recommendation and to set standards for them. What sort of questions should the regulator be asking the companies about the way in which they rank, select and amplify content on their platforms? What do we need to know about that process in order to regulate it effectively?

Renée DiResta: Thanks for inviting me to participate here. One of the challenges when we talk about algorithmic recommendation and curation is that we are talking about many different things across many different platforms, so having some clarity on the source of the concern is a good first step.

When we talk about recommendation and curation, there are two things that people most often refer to. The first is how content is ranked in their feed. For example, when you open your app, what order do you see posts in and how are things curated for you? If you use YouTube, what might be coming up in auto-play? It is about the ways in which information makes its way into your personal feed.

The other piece is a little separate, which is recommendations that encourage you to join particular groups or follow particular users. That is not part of the feed curation process. That is something that is suggesting that you expand your online behaviour through joining a particular community or following additional people, so a bit of a different dynamic is happening there.

There are additional algorithms that surface and curate things like trending, which are sometimes personalised to you and oftentimes also include an element of consistency throughout a geographic region—for example, things that are trending in the city of New York. That, too, is a slightly different product surface in which algorithms are used to surface,

in aggregate, something that lots of people are paying attention to. There are a range of different types of curation, recommendation and ranking algorithms that go into these products, often with multiple types of them on the same platform.

The Chair: From the work that you have done at the Stanford Observatory, do you think that data about people's location is important in this? If, say, you live in a community that is very strongly Republican, are you more likely to see pro-Republican information in your newsfeed because the systems are detecting that people who are in close proximity to you are largely interested in this? It is not based just on what you do but on the people you are physically close to.

Renée DiResta: It depends. A lot of times, yes, location and geography are incorporated. You can choose to turn off particular features. You can choose on Twitter not to disclose where you are. You have a little bit of control over the extent to which you want that incorporated into your feed.

If you create a Twitter account in a new place, oftentimes the first accounts that will be suggested to you are, for example, the local sports team. This is seen as a way to pull you into conversations in communities that you are geographically part of, not for nefarious purposes but because it creates a good user experience and a good onboarding experience for people who are new to the platform and part of that region. There are a range of reasons why that would happen.

In terms of recommendations, you will often see things that are geographically relevant to you. If I were to go to Facebook, buy/sell/trade is a big dynamic, like an online classified system. If I were to search for that, it would return results to me that are tied to the location that I am searching from or living in.

The Chair: Guillaume, you worked on the development of the YouTube recommendation algorithm. What questions should the regulator be asking companies like YouTube about the data that goes into that and the recommendations it makes?

Guillaume Chaslot: Thank you for having me. YouTube recommendations account for more than 70% of all views on YouTube, so it is really important for regulators to know what content is most amplified by YouTube. YouTube has this content and gives it to channel owners, so that they know exactly how many times they have been promoted by the algorithm. Regulators need to have a sense of that, at least at the channel level and at the topic level, in terms of the recommendations. The regulator should ask which content is already available to channel owners.

The Chair: What sort of data does the regulator need access to? Renée mentioned this earlier. We often talk generally about the recommendation algorithm, but there are many algorithms. How does a regulator go about looking at the different components that lead to

recommendations being made, or should it just focus on the content that has been recommended and not the system?

Guillaume Chaslot: It should just focus on the content. Knowing what the code is is not really important. What matters is what it does in the end. The code will be intellectual property. For me, it is about focusing on the content. The metrics are already present in the content that is given to channel owners, so we can just ask that. We should have that especially for content that is removed from YouTube. For instance, 0.17% of all content viewed on YouTube is removed. That seems little, but that is a billion hours a year. How many recommendations did this content get? Did it get tens of billions of recommendations? We need to know that.

The Chair: Is that information that the company has readily available if it chooses to share it?

Guillaume Chaslot: Yes, the company has it for channel owners, so it should share it. For me, it should share the maximum amount of information publicly. If there is a good reason not to share it publicly, it should be shared with regulators.

The Chair: In general terms, the problem is that, as we have discussed in the context of Facebook, the recommendation tools are principally there to hold engagement. "Up next" is there to keep you watching for as long as possible. Do you think there is a deficit of concern about user safety as long as engagement is high?

Guillaume Chaslot: Yes, exactly. You have cases where engagement is good for the user. When I listen to music, the longer I listen, the better it is for me. When there was a problem with paedophile content on YouTube, they spent a lot of time on the platform, so the algorithm was trying to maximise the amount of paedophile content that was shown to users.

Q93 **The Chair:** Laura Edelson, a lot of your research work at New York University has been focused on advertising data on Facebook, particularly political advertising. There is a question for the committee about whether advertising is in scope of the role of the regulator. At the moment, it is not, but that is something that we have to consider in our recommendations to the Government.

What sort of data do you think people should have access to when it comes to why they are being targeted with ads and why they are receiving them? Perhaps you could say, from your research, a little about what you think feeds that process. Is it just people's known political opinions, or is it other things about them that are used to target them with ads?

Laura Edelson: First, thank you so much for inviting me to give testimony today. There are two points that are really important here and which I do not think have been brought up yet. First, content and groups are often recommended on the basis of edge effects. This is not what you

necessarily have shown interest in but what your friends or other people in your social graph have shown interest in. Getting clarity on this point could be extremely helpful, because one of the things that we are hoping to do is to better understand how harmful content is recommended and promoted, so understanding these edge effects could be very helpful.

The other thing is bigger but often just does not get discussed. The single biggest thing that goes into content promotion algorithms that we know about is user interactions. These are things that are really visible to the user, such as likes or other kinds of reactions, comments or shares of content and that kind of thing, but it is also information about how the user has interacted with similar content in the past. For example, Facebook knows how often you watch videos and how long you watch them for. TikTok really masters this and really understands what content you will swipe and what you will sit there and watch. This user interaction data is incredibly important, and getting better visibility into this would, again, help us to better understand how to slow down the most harmful content.

I would say that those things apply, generally speaking, both to organic content and to ads. The border between paid and organic content is really porous, and almost all these things apply whether we are talking about an organic piece of content or an ad.

The Chair: What was the response you had from people who took part in your research project and who willingly wanted to understand more about the data that was being used to profile them to receive certain ads? They have no control over what groups they are included in, and if they are targeted through lookalike audiences, there is no knowledge and no consent as to why they are receiving it. Do you think this is information that people want, and were you disappointed that Facebook decided to stop you doing it?

Laura Edelson: I very much think this is information that people want. It makes people profoundly uncomfortable that Facebook is coming to conclusions about them that they do not have any input into and that might be completely incorrect. One of the things that our project does for our users is give them better visibility into how they are being targeted. We give them, "Here is one panel with ads you have seen and how you were targeted for them", and we have certainly received many emails from users who say that this gives them if not control then a better sense of awareness about how they are being targeted by advertisers and by the platforms.

Q94 **Debbie Abrahams MP:** Good morning, everybody. Thank you so much for joining us today. I just want to focus on what the social media platforms are currently doing to understand the harms that they are causing. We know, through the testimony to Congress, about Facebook's research into Instagram and the impacts on the mental health of young people. We heard yesterday, in another evidence session, that there is still a paucity of data about the range of legal but harmful content, but also about some of the technology that is adding to that. How much do

you think we know, and how much do the social media companies know, about the range of harms that they are causing?

Renée DiResta: We do not have very much information as outside researchers. The constant, many-years-long gripe is that we do not have very much visibility into it. We do not have visibility into what is recommended and to whom. We do not have much visibility into the impact that things have. We might see that content appears to be getting a lot of engagement. It is often difficult to necessarily translate engagement data to impact. We do not have impressions data a lot of the time, so again engagement becomes the proxy or the best thing that we have to work with on the outside.

Within the platforms, as has been evidenced by things like whistleblower testimony, there is, of course, more research going on into dynamics that are taking shape on the platform. That is the sort of thing that my colleague at Stanford, Nate Persily, is broaching legislation in the US for, to work on that problem and to address the problem of how we, as outside researchers, can make informed claims about these things, absent that kind of visibility.

Guillaume Chaslot: When I talked to people inside the companies, I was surprised that they did not really have any insights themselves. They were not proactively researching the possible harms. They were responding to public pressure. When people spoke about harms like those on YouTube, they responded, but they explicitly told me that they were not proactively researching those harms. I believe that they have no or very little incentive to do so.

Debbie Abrahams MP: Guillaume, can I ask for your views about the safety by design principles? We heard yesterday about how this may be a step forward for companies. If companies are not doing their own research, is this a way to help them in the process of becoming more proactively safer in what they are doing?

Guillaume Chaslot: Yes, that is definitely helpful. Especially in this Bill, we are forcing companies to do risk assessments, which is really good and is even better than what was there, although they will have no incentive to do very in-depth risk assessments. We should not rely only on the companies to do the risk assessments, but on the regulators and on civil society and researchers like Renée.

Laura Edelson: I just want to echo what Renée and Guillaume have said about the importance of better transparency overall, but we really know very little.

The one other thing I would like to stress is that one of the problems in even discussing these areas is that there are layers of lack of transparency. For example, all the companies that would be in this category 1 platform definition have policies against harmful content of various kinds, but what those policies mean in practice differs between platforms, and it is often not really clear where the line is when content is

too violent or crosses some other threshold where it might be threatening to another user. That is another common policy.

We just do not have a clear understanding of what these policies mean in practice. Users are often surprised and say, "The platform has a policy against threatening content. This piece of content was threatening. I reported it and nothing was done". Was nothing done because it does not meet the platform's threshold for that kind of content, or because it was an error in moderation? This is something that we just do not know. There is the policy transparency aspect that needs to be made much clearer, and then there is the enforcement transparency, where we just need much more information about how companies enforce their own policy.

There is a last piece. We know that Facebook does some of this, because it has started to report on it. It would be really helpful to know whether companies are evaluating—frankly, they should be—their own enforcement decisions to determine, in an auditing way, how accurate their enforcement is. What is your policy for enforcement, how much do you enforce and how accurate is that enforcement? We really need to understand all three of those things.

Debbie Abrahams MP: Is there something about definitions as well, Laura?

Laura Edelson: Yes, that first piece of understanding about what those platforms mean in practice is a definitional question. Sometimes, these policies are just vague and could be better written, which, in and of itself, would be helpful. Until you can understand what those words mean when they are applied to content on a platform, those definitions do not quite have meaning until you see what the platform thinks that they mean.

Q95 **Lord Clement-Jones:** Following up on the algorithm questions, this question is directed at you first, Guillaume. We have heard quite a lot about the function of algorithms and the approach of platforms to the deployment of these algorithms, but we see a paradox whereby harmful content and disinformation make up a small proportion of the content but large numbers of people report seeing it. In what way is this related to the way algorithmic systems govern content promotion and curation?

Guillaume Chaslot: That is a great question. The problem is that these algorithms create filter bubbles, where some people get to see the same type of content all the time. Even if a small proportion of the population sees disinformation about a topic, these people could see disinformation over and over again. A smaller proportion of the population gets very disinformed. We do not realise it, because people who are not in this bubble do not realise it. Thanks to research by people like Renée and Laura, we get more insight into these bubbles.

Lord Clement-Jones: I will come to Renée and Laura in a minute, but would you say that that is a by-product of the way the algorithms work, or is this part of the training of the algorithm as part of the business

model of the platform?

Guillaume Chaslot: It is a by-product of the way they work and the business model, so it is both.

Lord Clement-Jones: It is both, so if they were audited or evaluated properly by the platform, that aspect could be eliminated.

Guillaume Chaslot: If the business model changed, there would be different issues, depending on the business model. We always have to go back to the business model, because algorithms will amplify content that best matches the business model.

Lord Clement-Jones: That is helpful. Thank you.

Renée DiResta: There is one other thing that I want to introduce into your question here. Laura alluded to this as well. There is content that is recommended because of indicated interest in the past.

Lord Clement-Jones: You are breaking up a little bit.

Renée DiResta: I am sorry. I am on a hotel wi-fi. Then there is content that is called collaborative filtering, which is based in part on your peer group, your filter bubble or the communities that you are part of. If you are part of a community that has an interest in conservative politics, you will likely see more things that are related to that topic because your friends engage with them as well.

There is one other piece of this, which we call affordances at Stanford: the process by which your friends share content. That human involvement in the virality process is also a component of what is happening in the scenario that you set up in the question, which is: why do you see certain things and is that is tied into a particular filter bubble? If you have a finite number of friends and you are seeing posts from them, you are also going to be seeing the things that your friends have chosen to share.

One of the things that we see in the dynamics in the US—we saw it, for example, in the way narratives related to our election traversed the platforms—is a demand function, where users see the content that resonates with their pre-existing beliefs. They share it and it hits their friends' feeds. Their friends feel compelled to share it, so there is that dynamic that is happening as well.

It is not quite the same thing as the automated recommendation process. It is moving through an established peer group. It filters into the curation process of the available posts that the platform can show you. Which one is it going to show you? It can decide which of your friends' posts to show you, and you will likely see the posts that are getting a lot of engagement because, as Guillaume notes, the business model is that they want to keep you engaged and on-site, so they will pull from this array of posts that they have available to show to you.

That is not solely a recommendation function in the sense of suggesting something that you do not already engage with. This is something that is curating from the posts that are shared by your peers who are in your community, because ostensibly you want to see the things that your friends are talking about. It is a bit of a different process there.

Lord Clement-Jones: It is partly a fundamental network effect and partly the curation, but how do you distinguish between the two when you are auditing the operation of an algorithm?

Renée DiResta: I cannot really answer that as an outside researcher with limited visibility into what selection criteria are specifically being used. I do not know whether Laura wants to tackle that one.

Lord Clement-Jones: Laura, could you give me your take on this?

Laura Edelson: Yes. First, I want to very gently disagree with Renée a little here. In any category 1 platform that we are talking about that I know of—admittedly, I am an independent researcher—outside of email, there is no action that a user can take in the public news feed or in a public Twitter feed that will guarantee that another user will see that piece of content. Every action that you take in an interaction only feeds into the likelihood that a recommendation algorithm will then show that to another user. If you click a "share" button on a piece of content on Facebook or you retweet a tweet, that makes it more likely that your friends will see it but it is not a guarantee. These edge effects certainly matter, but fundamentally I believe it all comes down to the content recommendation algorithms. That is one thing.

There is something else that Guillaume said that I wanted to build on. My research has focused on Facebook, so what I will say I know because of specific research I have done on Facebook, but I have no reason to think it would not apply elsewhere. Misinformation is more engaging than factual content. It just is. The effect in the United States is distinct from any effect of partisanship. That is to say that both far-left and far-right misinformation are more engaging than factual content, and the effect overall is large. In US news, during the time period in which I did this study on Facebook, it was a factor of six; misinformation was six times more engaging than factual content. As I said, I did the study on Facebook in the US on US news, but I have no reason to think that I would not find similar things elsewhere.

The problem is that engagement and user attention are what these platforms sell. They sell advertising. This is their business model—to engage users. They have built algorithms to do that, and at least certain types of harmful content are just more engaging. This is the root of the problem that we are all trying to deal with.

As to whether these algorithms could be built not to promote harmful content, they certainly could, and we have seen this in some of the internal research documents that have been leaked. Those researchers were considering other factors such as reputation of the speaker or some

understanding of the content itself, so there certainly are things that could be done, but it all comes down to this central problem that harmful content can be extremely engaging, and engagement is what these companies sell.

Lord Clement-Jones: We have to look at the business model. It is a fundamental issue.

Guillaume Chaslot: Wikipedia, for instance, is a collaborative site where people could spread misinformation very easily. The structure of the site is designed not for engagement but to converge to more truthful information. It is easier to remove than to add misinformation on Wikipedia. That is why Wikipedia has a higher quality, on average, than Facebook, because the business model is different. That is what we have to study and then see that this business model is amplified with algorithms. We need to find a way to maximise this business model.

Q96 **Darren Jones MP:** We might like to live in a world where targeted advertising is not a permitted business model, but this Bill is not going to change that. I am interested in how the provisions of this Bill are operationalised within the technology companies and what effects the provisions of the Bill will have on outcomes.

We have talked a lot this morning about the level of automation on all these platforms. I do not know if anyone has ever added up how many individual algorithms there are that make these platforms work, but it sounds like a lot. The Bill asks companies to assess and audit their algorithms and to report on certain outcomes to the regulator. I then think, "What does that mean if you are at one of these companies?"

There will be people, presumably, who write the algorithms, and there must be lots of people if there are lots of algorithms. They must know how their algorithms work, unless they are fancy machine-learning algorithms, and then maybe they do not. There is then the product team that builds the products driven by the algorithms and will know to a certain extent how much these algorithms work.

You then have people who, like me in my previous life, live in the legal/regulatory/compliance team and whose job it is to do this work, and who have a meeting with the product team and say, "Explain to me how this works so I can write this up, because I need to write this up". You write it up and submit it to the regulator, having mitigated the risk to the business and putting a position that you think is in good faith and honest but is presenting probably the minimum amount of information that you have to present based on the law.

Then you have the PR team that says to people like us, and me now in my current life, "Everything is fine, because we've done all this work and here is our report".

I am just a bit concerned that, even though this is no longer self-regulation, it puts a lot of the responsibility on the business to do this work and to report to a regulator, which then has to have appropriate

capacity to challenge those reports and to ask for information.

My question really is whether you agree with that assessment. How difficult is it going to be to get these companies to truthfully audit and report on what their algorithms are doing? Who in those businesses knows the answer to that question and therefore should be mandated by law to submit that information to the regulator? Laura, you seem to be laughing, hopefully because you agree with me, but I will ask you to go first.

The Chair: Everyone is laughing now.

Laura Edelson: I am also laughing, because I have never worked at a social media company, but I have worked inside large Silicon Valley companies and am certainly familiar with the situation you describe, albeit from the other end. I was the engineer writing the algorithm.

The focus needs to be not on algorithms but on their effects. It is frankly impossible to describe how an algorithm works in a way that would make sense to a general audience. Any algorithm of sufficient complexity is not one algorithm but multifaceted, and often one part of the algorithm will undo something that another part of it has already done. The focus should be on effect.

You asked a really important question, which is how we verify what the platforms tell us. How do we know that this assessment is, first, truthful and, secondly, accurate? Platforms could tell you something in honest good faith and just simply be mistaken. If you later found out that something was incorrect, how would you know which one it was?

This is why, in addition to assessments being focused on outcomes and on effects of algorithms, this really needs to be paired with data transparency. Otherwise, there is absolutely no way for anyone to look at that risk assessment and know what it means.

Darren Jones MP: That is really helpful. When they do their auditing, it is about reporting on the effects of the algorithms and then having the ability to validate those effects, based on the data that is shared with the regulator, and then maybe even having a duty on the companies to report, if, after their submission, they realise that something was wrong has changed. It is an ongoing duty to report if the effect is different from that which they have already reported.

Laura Edelson: Yes, and my expectation, frankly, is that these effects will change over time, because these algorithms are ever-evolving. That is one piece of this. You have to remember that an algorithm without the data going into it just does not mean anything.

One thing that I am thinking of is that my team used to monitor all political advertising on Facebook in the United States, Canada and most of western Europe. Watching, for example, Covid being politicised in the United States in early 2020 was a very strange process, because we watched the input change. Then there was a delay and we watched the

algorithm change. The ultimate effects can change without anyone change the algorithm, if the inputs change.

Darren Jones MP: That is really important, Laura, because here in the UK we have some regulatory overlap, where the Information Commissioner feels responsible for the input, the data, and Ofcom is supposed to be the regulator for this Bill, which is about the outcome or the output, so it is important that we have a view on that as a committee. Thank you. Guillaume and Renée, do you have anything further that you want to add?

Guillaume Chaslot: I completely agree with Laura that we need to look at the output, so we need to have some output that is completely public. There are already some transparency reports by the platforms, but there is not enough data in them and they should be bigger. There should also be data for regulators so that they can validate and check the outputs, because we cannot trust the companies to do that.

One thing that would also be great to look at, other than the output, is objective functions. An algorithm or an AI has what we call a loss function, so it is what the AI is optimising. For YouTube, for many years it was watch time. It was about maximising watch time. This loss function is very important to understand and it is usually closely linked to the business model.

Renée DiResta: I would agree with Laura on this topic. We may discuss later the question of transparency reports, which, right now, are largely summary statistics that do not provide very much visibility. As Laura knows, the question of what is reported and who validates it will be one of the key ones moving forward.

Darren Jones MP: Yes, we will come back to that later, thank you.

Q97 **The Chair:** Laura, you said in your research that misinformation spreads six times more effectively than reliable information. Is that information that companies gather? If a company is saying, "There are lots of reports saying that we have a problem with misinformation on the site and with anti-vax conspiracy theories", are the companies producing reports internally saying, "We can see that anti-vax conspiracy theories are reaching a wider audience and being more actively recommended. If we wanted to do something about that, we would need to make an adjustment, and this is the adjustment we would make"? Does this sort of information on what they understand about what is happening in their own business and how they are trying to moderate it exist in plain language within the companies, which a regulator could ask to see?

Laura Edelson: We know that at least Facebook does a good deal of internal research into some of these questions. I know, for example, because it has been leaked, that it has done research into how anti-vaccine content has spread on Facebook. We do not know much about other platforms. I am not aware of research at other platforms into how harmful content spreads. It may exist, but I simply do not know. It would

certainly be relevant to ask platforms, in the context of this Bill, what kind of research into the spread of harmful content they have done.

These numbers come from research that my team has done. Facebook at least certainly has not disputed those numbers. In general, as Guillaume mentioned earlier, platforms are really disincentivised from doing this work, so I suspect that they would not put as hard numbers on it as I have done.

The Chair: Guillaume, do you think that YouTube has this information?

Guillaume Chaslot: There have been reports that some people at YouTube tried to research, for instance, the alt-right but have been discouraged by the management from doing this research. They are too scared when they do this research. Maybe nowadays they are doing it, but back in the days they were too scared that doing this research would be leaked or might force them to do something. They have no incentive to do this research. If they do the research and find something, they have to change it, so only bad things can come out of it. It is very touchy for them.

You have to beware what you ask as a regulator, because if you force them to give you all the research they have, they have even less incentive to do research, because if they find something bad, they have to report it externally. It is a very tricky place being the regulator and thinking about the platforms' incentives all the time. They will follow their incentives.

The Chair: From what you are saying, it sounds like companies can research why misinformation and harmful content spreads. They could understand and mitigate it, but on the whole they do not even do that work because of a fear of what it would show.

Guillaume Chaslot: Yes.

Q98 **Baroness Kidron:** Good morning. Before I start, can I just thank you for all the work that you all do? It is incredible and we are grateful.

I want to ask some really practical questions. One is about classifiers and the role they play in us understanding how the platforms themselves understand what is going on. If you go into the whole area of IP, what is marked and what is classified, or into radical or extremist content, it is known, classified, hashed and found. In the middle, it seems that there is an awful lot of content that we are struggling with in this Bill that is in no way labelled or attempted to be classified. I just wonder whether any of the three of you has an idea of whether we should even be thinking about and looking at that.

Guillaume Chaslot: YouTube has improved a lot on terrorist content from when I talked to them in 2017 and told them I was seeing terrorist content promoted by YouTube. They acted on that and solved a lot of things with very radical content. Right now, when I see the output of the algorithm, I still notice that it is very divisive. It is pushing people

towards the extreme, not with completely radical content but with more extreme channels. Now, it could just be Fox News or MSNBC compared to CNN, but they are pushing towards filter bubbles, so we need to understand how much they are doing that. Right now, we have no data to understand that.

The best way to address this type of content is to ask for recommendation data, watch-time data and how many removed videos and views there were for each category. That is the best way to study this part into which we have no insight right now.

Laura Edelson: You are really asking about the fact that a lot of content falls into a grey area. We have some known areas that I feel we could sit here and talk about, such as the harmful but not illegal content; the thing that immediately comes to mind here is perhaps content promoting self-harm or eating disorders. However, there are certainly other grey zones and I am not going to pretend that there are not.

There are two things, though, that are really important to talk about here. The first is that many platforms have existing systems to identify certain areas of harmful content, but they are really not very transparent about what those are, often because they are still figuring that out themselves. This is another area where there are some dangers of creating perverse incentives, so I would be wary of too many requirements on platforms on those areas of harmful content. I would also note that this is another area where new areas of harmful content can develop very quickly. We want to incentivise platforms to try to make the information ecosystems that they control spaces that are as healthy as possible. That is one thing.

The other thing is that, in understanding these grey zones, these things exist in general public discourse as well, but we are able to work those out in public in our common social discourse, because, again, the general information space of you and your friends having a conversation is open and accessible to everyone who participates in it. That is just not true of online information spaces. We just do not have enough data. Ideally, we would develop a taxonomy of the variety of harmful content that spreads online and there would be research saying, "We have developed a classifier and it is X% effective at identifying self-harm content". Someone else would come out with a better one.

That is the normal process of scientific research, but we just do not have the data to do that, which means that every platform is left trying to solve this problem on its own. The ability to share information between platforms on these kinds of issues is extremely limited. This is another area where better data transparency would really help the public to have a better understanding of the information spaces on platforms, and would help platforms create healthy information spaces more easily because they could finally share data with each other.

Q99 **Baroness Kidron:** I will come to Renée in a minute, but I want to just pick up on what you said. I was worried in the earlier part of your

conversation that we could do something very perverse to make platforms share research in a way that makes them look bad, and then they continue to fail to do it because why do research that will make them look bad?

A lot of discussion that we have had and the evidence we get splits into how we frame independent research, given that a lot of research is not that independent, in a way that is useful to your community, and how we frame the necessary oversight regime with the regulator so that they can have some sort of market intelligence on this as well so that what they are looking at is the relationship between purposes, data, weighting and outcome—the whole system. That should then throw up more known issues in a less fractious, broader setting.

Laura Edelson: That is a great question. Again, I apologise for sounding like a broken record here, but this is why it comes down to data transparency. We do science. It is not our job to take anyone's word for it. I do not want to say that it is useless to take the platforms' research without seeing the data that backed it, but it does not advance science about what is going on in these platforms.

Although there certainly are corners of data, particularly private data, that cannot be made public, there is a tremendous amount of public data on category 1 platforms that could be made public without compromising user privacy or creating concerns on data on specific users, which would really help us to understand how content-promotion algorithms work, how harmful content spreads in public, and the interaction between the platform and content promotion algorithms, and the effects on users. We can do that without studying users at all.

The other benefit of this approach is that it does not create perverse incentives for platforms at all. If we just make public data on platforms available to researchers and journalists, and ideally the public, we can come to our own conclusions. We can go through the scientific process of understanding various areas of harmful content and how we can avoid promoting them. That will help to make those platforms healthier places for users and start to rebuild some of the public trust that, frankly, has been lost.

Baroness Kidron: We would be very interested in the framing of what that looks like. If you have any thoughts about that afterwards, it would be great to hear from you. Renée, do you have anything to add?

Renée DiResta: I too would just jump on the "We don't have access to data" bandwagon here.

Q100 **Baroness Kidron:** Guillaume, can I finish by asking you a very detailed question? In your very first answer, you said that YouTube recommendations result in 70% of all views. I do not know whether you have noticed recently that it has just turned off auto-play for under-18s as a result of the Children's Code. I just would like you to explain to me where you think some of these features that encourage certain

behaviours, like auto-play, intersect with the "recommend" algorithm. Is it the combination of those two things or is it just the algorithm on its own?

Guillaume Chaslot: That is a great question. It is definitely a combination of the two things, but I want to point out that these features are built by humans who have to think about the features and how they will affect users. The algorithm improves by itself, so it is always improving. It can get much smarter than a human, so I am more scared about the algorithm. In some ways, it is much smarter than me in finding what I am likely to click on. I am more scared of that.

Baroness Kidron: My colleague Debbie Abrahams asked about safety by design and we should be looking at those sorts of issues: the features, the push, as well as the algorithm. We must not lose sight of that. Thank you very much.

Q101 **The Chair:** On that, the bottom line would seem to be that the more vulnerable you are, the more likely you are to see harmful content. The greater an interest you take in content displaying self-harm, the more likely you are to see more of it. Would that be a correct assessment? Laura, you are nodding.

Laura Edelson: It is good that you asked about self-harm, because, while I do not know that we have public research to say this generally, for this one category some of Facebook's internal research about this point has been leaked to the public. Frankly, yes, this gets to the heart of interest-based content promotion. If you are interested in content on weight loss, you will be promoted more and more extreme content on weight loss. That is, generally speaking, how content promotion algorithms exist. If you have a casual interest in something that might be harmful, you will be promoted more content about that, which certainly has the possibility to feed into that harmful interest.

Guillaume Chaslot: It should be seen as someone who has an interest in self-harm possibly going through a bad period. The ideal algorithm when you are going through a bad period is one that tries to cheer you up, but the algorithms of many platforms are trying to keep you in this self-harm mentality when you have fallen into it. A friend of a friend fell into a rabbit hole of conspiracy theories. At the time, he was very vulnerable because his father had died. The tricky thing is that it does not give you what you want. It takes advantage of one vulnerable moment to try to recommend things that are harmful to you.

Laura Edelson: Just to follow up briefly on that really horrible specific example, content promotion algorithms very often work on what you interact with. User interaction is not me watching a video, sitting and thinking about it for five minutes and coming up with a measured reaction. User interaction is really quick. It comes from the part of your brain that is inherently reactive. It comes from the part of your brain that, from an evolutionary perspective, responds to threats. It is key in telling you that this is something you need to pay attention to.

Not enough research has been done on this point, but there are researchers who have done some preliminary work that indicates that certain categories of harmful content trigger that part of our brain that is key to identifying threats. You will certainly notice that content very quickly. Again, this is an area that needs more research. There is certainly the possibility that we not only notice that kind of content much more quickly than we notice puppy pictures, but we react to it much more quickly. Algorithms that are built on user interaction would pick up on that and would then promote more content like that, because users would be interacting with it more.

The Chair: I want to ask one final question on that. Renée, it goes to something we spoke about earlier, which is the sort of data that informs those recommendation decisions. User interaction with other content would be one area. We spoke about proximity data earlier. We know that companies like Facebook gather data via Bluetooth from the proximity of devices and people who you are with. Could that sort of data—GPS data or maybe data gathered from people using a common wi-fi router to go online—inform these selections as well?

In the context of self-harm, let us say that you have a young group of female students in a shared dorm at a university or college, and some of them are engaging in self-harm and some are not. Would all of them be more likely to see self-harm content as a consequence of that? The data that is driving that selection is not just data about their own interactions but data that is gathered about who they are with, what times they are with them, what they like together and, therefore, whether they are more likely to see more of the same thing.

I ask, because it would be dangerous and very disturbing if that was the case, and it could also make the case that certain datapoints may be collected and used in this way that should not be.

Renée DiResta: I am not sure about things like proximity or shared wi-fi. That is a little outside my understanding of the inputs into the algorithm. If they are in the same dorm, they are possibly friends with each other, so there is a much simpler way by which that would happen, which is that they are just friends. They share social connections on a social graph, so that could be enough to surface some various types of content.

What I was alluding to before on this difference between a recommendation versus curation is that you can recommend new accounts to follow based on interest. If you choose to follow self-harm or ED-related content, you may see more of them.

The other thing is that, once you are part of a community or have been recommended into a group and you choose to participate in a group or follow a user, what you will see is in part keyed off of the conscious decision or choice that you have made to follow that account, which the curation algorithm interprets as interest in the content. If you are part of a Facebook or Twitter group, or follow a particular channel or influencer,

you are opting in to seeing that influencer's content in your feed. For groups in particular, there is a feature where you can select to make sure that you see all posts, which means that at that point you have indicated an elevation in your interest in that community, ergo you will see more of it.

Additionally, if the group is high volume and people are potentially sharing a lot of information, which is very common right now in anti-vaccine or anti-mask mandate groups, which are highly active communities, you see more of it, again because these are groups with very high degrees of engagement. There is a bit of a different dynamic happening there at that point.

As for your question about how you might see things that your friend group sees, the simple fact that they are your friend group will likely be enough to precipitate that.

***Guillaume Chaslot*:** Like Renée said, it is based more on the interaction and the social network. To give an example of harmful content that I got recommended, a friend once shared a site with anti-vax information. I went on the post and the URL was not even right. I commented on the post, "You are sharing a false URL". Afterwards, Facebook thought I was interested in anti-vax content and started recommending anti-vax content directly to me—"Suggested for you"—not even through a friend. It was just bombarding me with anti-vax content because of this one comment. Maximising engagement is really stupid sometimes.

The Chair: Do you think it is the case, as we had put to us in a previous inquiry, that although we all think we are unique, people are actually quite similar? If you start doing something, such as engaging in anti-vaccine content, the algorithm kicks in at that point, because suddenly you have done something quite interesting and it rallies on that, rather than your other interests, which are quite similar to other people who are like you.

***Guillaume Chaslot*:** Yes. That is what I have noticed studying the YouTube algorithm. It does not try to find the perfect content for you. It tries to find the rabbit hole that is closest to your interests. That is what I have seen.

Baroness Kidron: I am sorry to come back on this, but I just want to understand something. Historically, some of the platforms have told me that they are looking at the engagement, not the content. It is back to this whole grey area thing. The Bill is predicated in part on them upholding their own terms and conditions. I accept the earlier comments about us sometimes being disappointed about what their terms and conditions actually mean in reality, but let us assume for a second that we are talking about something that they do not allow but then they are then recommending because of the engagement. Is there a way of making them responsible for what they recommend if it against their own terms? At the moment, that is the Bill we have in front of us.

***Laura Edelson*:** There is no reason why not. Again, you just need a means to actually evaluate their compliance, which is the primary thing that you do not have in the Bill as it stands. Right now, you would have to take their word for it that their risk assessment is accurate and complete.

We have this problem with the FTC in the United States. There is a tremendous amount of content in advertising that does not comply with existing US law. It is fraudulent; there are scams. The FTC does not have a mechanism to easily get at that ad data, so it does not have a mechanism to enforce existing law. Asking platforms to comply with their own stated policies is entirely reasonable, but right now you do not necessarily have a mechanism to know when they are and are not doing that.

Baroness Kidron: This is the oversight transparency bit being absolutely mandated and in the hands of the regulator rather than in the hands of the platform. For the record, I am seeing some nods. Thank you.

Q102 **Dean Russell MP:** I would quite like to explore a little more the data input part and how sophisticated that is. As with any system, with bad data in you get bad data out, or in this instance bad recommendations. I was just interested to understand where all those data sources are coming from for the different platforms, because we often talk about Facebook, YouTube and these channels as very individual, siloed systems, but, as we know, they talk to each other in various ways; you have browsers talking through cookies and all these different things happening.

I would just be interested to know from you whether we would be shocked at the scale of the data that is collected about us and used to go into these algorithms to make these recommendations. What sort of sources are out there that are being used currently? One of the conspiracy theories you often see online is that Alexa might be hearing you and listening to conversations. The Chair mentioned earlier Bluetooth proximity to other people. There are all these things. Would you mind explaining your views on that data input piece, please?

***Laura Edelson*:** I know there are absolutely conspiracy theories about Alexa. Those do not hold a lot of water. As always, the truth is more banal.

When you use a platform such as Facebook, TikTok or YouTube, as you use that platform you are giving the platform a tremendous amount of data about what you interact and engage with and how you do it. Every move you make on a platform is monitored, because they are trying to optimise for your engagement. They want you to be as engaged as possible, for as long as possible, as often as possible. That is what the platforms we are discussing are built to do. You give them so much information about how you do it. How fast you scroll and how long you watch videos can be monitored. When you are on YouTube, which video are you likely to go to? That will tell them where your eye tracks to. They

do not care about where your eyes track to; they just want to get you to watch as many videos as possible for as long as possible. When they show you an ad, they want to know which ads you will click on.

As you interact with the platform, that data is tremendously powerful because they are always getting more data. As they make little tweaks, they see how you respond to those little tweaks, and then they can tweak further. This is used in a lot of different ways. This is used to choose which ads they will show you, which version of an ad they will show you, what content they will promote and what groups they will promote.

Damian asked earlier about the effects of proximity. That is a piece of it, but the proximity of your social graph and the edge effects are also tremendously powerful. This can mean that if you are in a social group that has been exposed to extremism, it will spread through that whole group because of those edge effects and because of the interactions of other people in your group. Even if they do not like it, if they interact with it and enough people interact with it, you will be shown it as well.

Dean Russell MP: Is that just from that platform, or are they collecting similar data from other platforms? If I am on Facebook, it obviously will know everything that I have done on Facebook and social graphs of friends and connections, but is it also pulling that data, or different types of data, from other places as well?

***Laura Edelson*:** To my knowledge, no. They have really strong business incentives to protect their data from other platforms, because it is the core of their business. They want to know as much as possible about their users and how their users interact, and they do not want their competitors to know that.

Dean Russell MP: Is that vice versa then? Will other platforms and browsers and so on not be picked up in Facebook data? If I only used Facebook and nothing else, then of course they would only have my data from Facebook, but if I use Facebook and lots of other websites, browsers and channels, are they factoring any of that into it, or is it always just based, in the Facebook example, just on what Facebook has from me on its platform?

***Laura Edelson*:** It is not solely based on Facebook, because you will interact with those platforms together. It is common that people post YouTube links to Facebook. When you do that, Facebook knows about what you are doing on YouTube. That heavily factors into what else they will promote to you on Facebook. Users themselves leak data cross-platform all the time.

Dean Russell MP: That is what I was trying to understand. That is really helpful. I have a quick point on your narrative just then. I must admit that the way you described the way that platforms optimise and optimise sounded a little like the way a drug dealer might make their drugs better to make a better high, constantly refining to give that saccharine effect or that high in a quicker and more impactful way, to get people more and

more hooked. Would that be a fair analogy, from your experience?

***Laura Edelson*:** Yes. As Guillaume put it earlier, these are just optimisation functions. They have specific things they are trying to maximise, which usually comes down to engagement, session length and number of sessions per week. They will continually optimise for those things.

Dean Russell MP: In a way, for young people and children, I will not go as far as saying it is rewiring their brains, but there will be a vice-versa effect there, will there not?

***Laura Edelson*:** I am not a psychologist. I have not done that kind of research, but that is certainly a reasonable conclusion to draw.

Dean Russell MP: Renée, I have the same question for you. I was interested in your views on the interconnectivity of data and how the platforms are using the data inputs and the sophistication of that, from your experience.

***Renée DiResta*:** Laura described it as I would have also. You mentioned browsing behaviour and open web. Certain websites use particular comment plug-ins, perhaps, from Facebook and other things. There is some visibility into what you are doing elsewhere on the web if that website has chosen to use particular facets that connect with Facebook in some way. Otherwise, it is this sharing across platforms of people posting a YouTube or Rumble link, or another platform link, from one place to another.

Dean Russell MP: I know that nowadays we have far more privacy options. I know that Apple is very much trying to lead on this, with the idea that you know what data is being shared from apps, for example. Does there need to be better protection for children on this? I would imagine most children do not know that they are sharing all this data online. One of the aspects of this Bill is protecting children, and one of those aspects of course is the recommendations, but we can limit those recommendations by limiting the data that platforms are able to use about children. Would that be a fair thing to look at?

***Renée DiResta*:** I believe so. I do not work on harm related to children, so I do not have very much knowledge of the psychological research there, but it seems like the sort of thing that you would be able to do.

Dean Russell MP: Mr Chaslot, on the same theme of the data sources, especially given that you mentioned experience with YouTube and so on earlier, is the data collected from people watching videos, for example, based solely on the length of time they watch them, or does the content from those videos, or even from images, get taken into account? If you watch a video for 10 minutes and the title is harmless—it might be about having fun on the beach or something—but actually the content shown and the language used is harmful, would that also be seen as a data source that could be used, or is it based more on the length of time, the

type of video and so on that has been put in by the creator?

***Guillaume Chaslot*:** A lot of data sources are put into the algorithm, but eventually you will usually have one data source that is more powerful because it contains more powerful information. That is the amount of time that you watch, in the case of YouTube and TikTok.

One reason why TikTok became so viral and so addictive with kids is that videos are shorter. The algorithm has more data on children than on YouTube, because the duration of a video is much shorter on TikTok. That is why TikTok is learning extremely fast from children. I am really worried about the amount of information that the algorithm on TikTok can get from these watch times. It can be used for what type of colour, music or body shape they like. It is very creepy when you think of everything that can be used from these interactions, and it is creepy that they can be kept for so long.

Dean Russell MP: We mentioned earlier platforms not using reports. It almost feels like, "See no evil, hear no evil, report no evil and therefore assume that it doesn't happen". If you take TikTok as a platform, you hear numerous times, on a weekly basis, about a new dare that is going on, with kids in schools being dared to do something; we have seen awful situations with the kids swallowing capsules for washing machines and so on.

Do you think the take-up of those is directly related to the content that is being gleaned? If somebody has perhaps done a previous dare and that is in the data, and they have shared that content to show that they have done that, say on TikTok, are they more likely to be pushed to do the next big thing that is coming through—the next dare, the next activity—that could be dangerous for their health?

***Guillaume Chaslot*:** Yes. There can be such a rabbit hole, with the kid starting to watch something, leading to them falling into more and more extreme dares.

Q103 **Suzanne Webb MP:** I concur with what Baroness Kidron said: thank you for all the work you do on this matter. I am very focused on people's expectations of the Bill. Guillaume, I will go back to the very first thing you said. I am pretty sure you said that billions of hours of content will need to be taken down. Did I hear that correctly?

***Guillaume Chaslot*:** About a billion hours of content got removed from the platform by YouTube. People have watched it for one billion hours and then it got removed. One billion hours of terrorist content would cause a lot of harm. We do not whether these billion hours per year are spent on terrorist content or just news. It is things that even the platform knew were harmful, because they removed them, but we have no clue what they were.

Suzanne Webb MP: I slightly misheard. How long will it take to take this content down? My biggest concern is that we can have the best and most robust Bill in the world, but how realistic is the timeframe to take it

down? We have all talked about vulnerable people. All this content is focused on vulnerable people, basically. We need to get this down soonest, at the earliest opportunity. How will that work? How quickly will this content be taken down?

***Guillaume Chaslot*:** It does not really matter how quickly it is taken down, because if the algorithm is amplifying this content a lot, even if it stays for one day it will get many views. One video got taken down and in one day it got 6 million views, because the algorithm promoted it a lot for one day and it got 6 million views. It was about Logan Paul, who went to the suicide forest in Japan and poked at dead people. That was pretty disgusting, but YouTube pushed it towards 6 million views, if I remember correctly.

It does not matter how long it stays on the platform. What matters is how much the algorithm is amplifying it. We talk about the fine that the platform will get, but if the platform is fined not by the time the content stays up but by the number of views it gets from recommendations, for instance, it will be more reactive and will try to not recommend bad content as much.

Suzanne Webb MP: I think I understand. I keep explaining that I am bit of a luddite when it comes to algorithms and things like that. I am more of a process person. My expectation would be that from day one we are focused on getting this content down as soon as possible. Does the Bill deliver that? Will the tech companies and regulators ensure that that happens?

***Guillaume Chaslot*:** The devil is in the details. If the sanctions are too strong, companies will be incentivised to delete a lot of videos and to do too much content moderation, which will be harmful to free speech. If the sanctions are too little, platforms will not make much more effort than now. It all comes down to how proportionate the response of the fine is to the harms that actually happen on the platforms.

***Laura Edelson*:** Just to bolster what Guillaume was talking about, there is so much content on all these platforms that it is important that the focus is on what algorithms promote and how much engagement content has, because there is content that exists for months or years and no one ever see it, but content that is up for a day and gets a tremendous amount of engagement. Changing the incentives for what content gets promoted is the most important thing.

Q104 **Darren Jones MP:** I have a follow-up question. I am keen to understand how nimble the platforms can be, whether through algorithmic or human content moderation. In this Bill, for example, the Secretary of State has the power to tell a regulator that some type of content is harmful and that they expect the companies to deal with it. We will put to one side whether that process is as it should be. Say that happens and the Government tell us that they need to do that because they want to respond to things quickly. How does the tech company then respond to that? If they say, "A new type of content has been alerted by the

regulator as being harmful and we want you to something about it", what happens on the business side when you receive a notice like that?

***Guillaume Chaslot*:** It seems that this process would be too slow in any case, because, as a regulator, by the time you notice that content is harmful and has an impact on a lot of people, it has already impacted a lot of people. This should be for only very extreme, very corner cases that you want to push down, but that mechanism will miss 95% of the problems. It will be too slow, because by the time it goes one way or the other way, the content has already been seen. You would need to do a counter-content, which would take too much effort, in addition to deleting the content. That is one thing, but it should not be the only thing you do.

Darren Jones MP: I understand. It is useful to know that the quick method is not quick enough. How does the tech company then deal with these issues? Say, for example, there has been an awful tragedy in the UK, a terrorist attack or something, and someone has put the video online. If the Secretary of State telling the regulator to tell the tech company to take it down is too slow, what does the tech company do? How quickly can you spot these things if they are not the well-categorised, well-tagged, well-trained bits of content that already go through standard moderation?

***Guillaume Chaslot*:** If there were penalties based on the number of views that videos get before they get taken down, the platforms would have an incentive to react quickly by themselves, without the need for the regulator, especially if they have amplified the content rather than just left the content up. We have to be really careful when we give incentives to the platforms to delete content, because it will harm free speech. Rather than telling them to delete the videos, just telling them to stop recommending it is much better for free speech, and it can be nearly as efficient in most cases.

Darren Jones MP: It is about amplification, not about taking the content down. That is useful. I still do not know how, in practice, this is flagged internally. Presumably people are monitoring what is going on the platform and what is doing really well in real time. Are there people with lots of screens looking at this stuff? How do they know that something has been posted on their platform and suddenly 6 million people have seen it? Do they have to wait for someone to report it or do they know? How does it work?

***Laura Edelson*:** There are a few mechanisms. Every platform that I know of has some mechanism to detect illegal content, because this is a problem they all deal with. There is content that is just illegal and they have filters in place to automatically detect that illegal content. For certain categories, they have invested a lot of effort into making sure that these filters perform very well, and they do.

You gave the example of a terrorist attack. If there was a specific video of a terrorist attack, they could filter that fairly effectively pretty quickly.

Renée, would you say that 12 hours is a reasonable expectation? That seems reasonable to me.

***Renée DiResta*:** I am thinking of Christchurch, which was an interesting experience.

***Laura Edelson*:** Yes. They did not do a good job on that. They could have done much better.

***Renée DiResta*:** Christchurch was the worst-case scenario. I was working at a different company. I was not at Stanford then, but we saw the post on 4chan go up and then we saw the video. It was unlisted on YouTube. It was posted privately first, and my response was to flag it for contacts at YouTube and Twitter, because he had posted his manifesto in tweet form. That was the first flag as an outsider, saying, "We see this. We don't know with 100% certainty what it is. We don't know if it is old footage being recycled. We don't know if it is new. We think it is real. It doesn't look like a hoax. It's tied to this post on 4chan".

We assembled, effectively, a quick list of evidence and then sent it to them. It took about 20 minutes for us to do that. At that point, this was beginning to go viral through the chan ecosystem, which was then saying, "We have to get the video down. We have a mirror that has put it up in a bunch of different places". From an outsider perspective—again, I cannot really speak to what was happening internally—we started to see selective cuts and different clips of the video, and ways to try to deliberately evade the hashes.

Again, the hashes were largely in place. Certain types of content were classified as terrorism and certain types of things were quickly hashed and dealt with. Through the Global Internet Forum to Counter Terrorism, certain shared repositories existed for certain types of content, such that when it hit one platform, other platforms could have access to that hash, find the video and take it down on their platform. There were these attempts, in certain content types, to create that kind of sharing ability so that it could be taken down as quickly as possible, but there were also these very adversarial dynamics whereby the people wanted to continue to try to upload it. That dynamic was happening as well. It was not quite so simple as taking down one video. It became a massive effort.

I cannot really speak to the dynamics internally. I know that it was routed on a few different teams. The people that we might report to are not necessarily the people responsible for it. It was actually dinnertime, Pacific time if I recall; it was 5.30 or 6.00. There was also that interesting gap of who responds to an emergent catastrophe, from a human capital standpoint. There were a lot of complexities that went into that.

There are increasing attempts to have things such as trusted flaggers or other outsiders who can flag content, indicating that perhaps a platform should look at it, because of things like a velocity change in a particular type of content. Again, there are certain areas where there is a much higher standard of care, such as terrorism and videos in which there is

some sort of human tragedy. There are particular ways in which they try to find these things and take them down as quickly as possible.

There are other things, such as routine misinformation and disinformation or hoax videos, that are not really treated in the same way and do not have that same rubric of sharing, and there are a variety of reasons for that.

Darren Jones MP: I understand if there is an original piece of content that you can categorise, hash and know when that content is being shared or played elsewhere. I understand that there are types of images or video content, whether it is child sexual exploitation or other types of content, where you can, by what is in the content, figure out what that is and filter that or deal with it. I am interested in this idea of a flexible or nimble response to something that is new as opposed to something that you can train your algorithms to expect.

For example, there was a lot of racism after a football match here in the UK a few months ago, and the Secretary of State said, "Under this Bill, we will make sure that that content isn't available". Presumably he meant—I can only assume—that, by this provision of telling the regulator to tell the platforms that he does not want that content on there anymore, something would be done. How on earth do you respond to that? That is different types of content with different contextual subtlety. Can it actually work in practice?

***Laura Edelson*:** We saw this play out in the US in 2020, and there are two cases that spring to mind. The first is one that I mentioned earlier, where we watched Covid and many of the effects of that become, frankly, politicised in the United States. We watched particularly Facebook begin to consider this a political event and begin to have some content policies on this and enforce those policies.

The other thing that springs to mind, again in the United States, is the summer of 2020 when there were a lot of protests on systemic racism. There was a lot of misinformation about those events, as well as extremists organising on those events. There was certainly an attempt by Facebook to classify and identify those content areas and to adjust to that in certain ways.

They can be very crude instruments. One thing that springs to mind is the mask shortage in the United States for quite a while in the spring and summer of 2020. Facebook had a policy against advertising masks online, because a lot of counterfeit masks were being sold and there were a lot of problems with that. That translated into ads with people wearing masks getting flagged and taken down, and there were a lot of ads of people in masks that were not selling masks.

Again, there is the capability to respond, but that response is not going to be perfect or highly accurate, especially in the early phase of responding to an event. Something that also may not be widely understood is that, in order to make a well-performing algorithm, you need a lot of data. Some

kind of new threat or new event emerges, and soon after, in the early period, just temporally you do not have a lot of data. No algorithm will perform very well at the beginning, but it will get better over time.

Q105 **Lord Gilbert of Panteg:** I want to explore a little further the issue that Darren has raised, which is unexpected events, what protocols platforms would have in place and who in the organisation would respond to those events. It may be that the unexpected event gives rise to something that is clearly harmful. On the other hand, there may also be a justification for leaving the content there. It may be in the public interest for the content to be left there, because it is reporting an important news story and disclosing facts that, however disturbing, the public have a right to know or because of free speech reasons.

Those are considerations that an algorithm or artificial intelligence cannot deal with. How does the evidence from the technology and the human decision-making come together? Is there a situation room? Who is sitting in that room, leading on behalf of the companies? Darren pointed out speed. Would there need to be some very fast-track engagement with the regulator in those circumstances? In your experience, does that exist in any jurisdiction?

***Renée DiResta*:** There are some experimental, early partnership efforts to think through how to do this. Our team at Stanford participated in one in the US 2020 election, where there were designated teams responsible for tracking election-related misinformation in the United States. They are usually called integrity teams. For other types of content, they are perhaps trust and safety teams. There are some teams focused on Covid. There are a variety of points of contact, and there are ways to create the trusted flagger relationship, where information is surfaced to people at the company whose job it is to then look at it. We are not, as outside researchers, telling them which of their policy enforcement rubrics to use. It is just, "This is a thing that we're seeing that appears to violate an aspect of your terms of service. Here is our awareness of the situation".

There are some arguments for, and I believe some platforms have begun to use, a circuit-breaker, which is a very positive potential intervention going forward. This is the idea that when content reaches a particular threshold of velocity or virality and appears to be snowballing and reaching a large number of people, you would want that to reach your fact-checker so that they can assess what is happening or to reach the relevant teams within the platform so that they can assess what is happening, and potentially throttle its distribution while that is happening if it falls into a particular type of content that has the potential to harm, as you know. You do not want to be doing this for every random online meme or hoax or what have you that goes viral, but there are particular ways to tailor that type of potential action for particular types of content where, in the past, there have been demonstrable harms.

Tying that notion of harm to interventions is key. As I think Guillaume mentioned, the general rubric for moderation has three types of interaction. There is remove, which is where it actually comes down, but

there is also reduce, which is what I am describing—the idea that you can temporarily limit its distribution in the feed as you take the time to determine what is happening.

The final piece is inform. You are potentially giving information to your fact-checkers or others who will produce the counter-content. In the case of election misinformation, that might be the relevant Secretary of State here in the US, or whoever is the appropriate counterparty in the UK. They would have the ability to say, "No. This is a rumour, and here is the fact", and so, again, to append some sort of information. Rather than taking it down and creating a whole second layer of conspiracy theory on the censorship dynamic, you are instead taking the opportunity to move into a content transmission environment that lends itself better to reduce or inform.

***Guillaume Chaslot*:** I completely agree with that. Reducing this type of content is most important, rather than deleting it.

Q106 **Lord Clement-Jones:** I wanted to follow up on that circuit-breaker point. Yesterday we were talking about safety by design. Should we incorporate the idea of a circuit-breaker that is triggered in certain circumstances into a code, as a requirement for design, for safety? Is that something we should actually prescribe up front?

***Laura Edelson*:** It is an interesting idea. You will forgive me for speculating here. This is where I would just love to have more data. I do not know what a reasonable threshold for viral content would be and how many pieces of content on a daily basis on one of these category 1 platforms meet that threshold.

It certainly seems entirely reasonable that content that falls into certain categories, gets significant engagement and goes viral should be reviewed to ensure that it meets platform policies at a certain point. That, on its face, is an appealing idea. I am uncertain what thresholds would be and how many pieces of content we would be talking about, because those things would inform how practical that would be.

***Renée DiResta*:** I would say the same thing. There is the question of how to legislate it versus advocating, as we do sometimes, for platforms to undertake it as a self-regulatory or research-type project. We are still at more of the latter phase at this point. I am not totally sure what the actual regulation would look like.

We see things in the financial market. The circuit-breaker model is actually derived from that idea. There are certain thresholds at which there is a significant amount of volatility, and a pause is given to gather more information and inform the public about what is happening. The idea derives from whether to do it in a regulatory model or whether it is something that platforms will do with design internally as a voluntary choice. That is an open question.

***Guillaume Chaslot*:** It depends too much on the details of the implementation. We have to be careful with the details of the type of implementation. It could be great or really bad.

Q107 **The Chair:** Before we close, I just wanted to pick up on a few things that were mentioned earlier. Renée, you mentioned transparency reporting, but listening to the evidence we have discussed today it would seem that that is one of the great misnomers, because there is no transparency at the moment, and the information that is produced by the companies in self-reporting is largely out of context and, to some extent, meaningless.

***Renée DiResta*:** I think I said that the transparency reports provide some visibility into the scale of the problem, perhaps. It is interesting to see the category breakdowns on particular things that come down. In the US, we have what are called the Santa Clara principles—the idea that there are certain ways to do transparency reporting well even if they are largely reporting summary stats. We do not have very much visibility into the kind of content underlying that, and that is one of the ways in which the transparency reports are not as valuable as they could be. They largely become just numbers.

One of the areas where you see this play out at times is that there will be an increase in something like the statistics on CSEM on platforms, where potentially an improvement is made in some form of detection, which then elevates numbers.

There are other areas in which there are adversary influenced metrics, such as the number of fake accounts on a platform, which may look large because an adversarial actor has come and really made a concerted effort to create a whole bunch of fake accounts, and then the platform actually takes them down and stops them, but the number increases, and that is then reported as if the problem has increased. It is not really servicing what percentage of fake accounts are coming down. It is just saying, "Whoa, this huge number of accounts has been taken down". There are just ways in which the statistics are not as useful as they could be in getting at the kinds of harms that you are interested in understanding.

The Chair: Following up on the discussion about a circuit-breaker effect on social media and how that could work, I was interested that one of the key takeaways in the *Long Fuse* report that you worked on, looking at misinformation in the 2020 US election, was about the role of influencers in driving disinformation. Could the regulator ask social media companies to pay closer attention to accounts, pages, channels and so on that have very large audience numbers, which are more likely to be super-spreaders of disinformation?

***Renée DiResta*:** In the US, that would get into some uncomfortable dynamics. Particularly since it is political content, I would not want to see a Government advocating that a tech platform pays more attention to a political party in opposition. There are some challenges on that.

The Chair: The *Long Fuse* report obviously focuses on the US election,

but we have heard this in other evidence on anti-vax campaigners: that there are a dozen or so super-spreaders of anti-vaccine conspiracy theories. If information is being distributed on a platform with very high audience numbers, the danger or the risk of content being spread in that forum is greater because it is far more likely to spread more quickly and be seen by more people. Given there are far fewer accounts like that that have very large audiences, or groups on Facebook that have large memberships, would it be reasonable for a regulator to say that you should spot more readily what happens on these larger platforms, because that is where the greater amount of harm could be done?

***Renée DiResta*:** These particular types of actors that you are describing are violating platform policy. The question becomes why there is not more enforcement of the platform policy as it exists, as opposed to, "Why have you not adequately dealt with these five people?" It becomes more a question of what the reasons are for these accounts to continue to persist.

One of the things that was interesting in the whistleblower testimony was the dynamics through which large accounts are in fact to some extent indemnified from certain types of enforcement. There is potentially a significant problem with that policy, in fact, but there are also ways in which that undermines their other policy, which is to reduce harmful content. That policy to reduce harmful content appears to confront this other policy that says that certain types of accounts are given greater leeway. That is potentially the sort of thing where we would want to see that dynamic reconciled in some way.

The Chair: It would be reasonable to assume that Facebook gave those large account exemptions because those large accounts drive engagement by their very nature.

Looking back at the 2020 election in America, you have been involved in work analysing what happened, particularly after the election date of 6 January. What sort of lessons should we learn from that? There is legislation in the US and the UK on incitement, which is a criminal offence, and cases can be brought against people for inciting others to commit violent acts, but when does that kick in on social media? Is it when the first stone is thrown? At what point should a social media company look at a sustained campaign of activity that could be likely to lead to imminent harm, and when should it intervene?

***Renée DiResta*:** There are policies related to incitement. Facebook took down some of the groups that moved in that direction. There were a number of very large groups on the "stop the steal" movement in the days after the election that came down. One real challenge, which I think will be dealt with very differently going forward, is this question of whether they are serious. That was a dynamic: that those who observed those posts in the US did not take them as seriously as they should have.

I do not want to speculate; we are still in the process of fact-finding on what was on what platform when, who saw it and who was responsible

for making decisions, particularly for the security at the Capitol and things, as a response to that. I believe that platforms have also suggested that they did try to flag certain aspects of this content for law enforcement and others in the US. I do not want to speculate on where exactly the ball was dropped.

One of the dynamics that we saw recurringly was very angry people forming groups. The real challenge is that this is how political activism manifests. There are people with real political opinions, exercising free expression. When it crosses that line and where that line is is the challenge. Having some sort of outside observation of those dynamics and understanding what is happening is an area where we need more visibility. Again, it comes back to the central theme that all three of us have emphasised, which is this question of access to the ability to understand these dynamics in as close to real time as possible.

The Chair: Do you think that line exists? Do you think that companies have tried to draw that line? It has been reported that Facebook decided on polling day in America to make News Feed reprioritise trusted sources rather than friends and family content, and then reversed that decision when it drove down engagement. The logic of that would be that, rather than there being a line that Facebook was policing, it was happy to see the anger because that drove engagement.

***Renée DiResta*:** On that question of what to service, what to curate and why Facebook shifted it in one direction and then changed it back, again, a lot of what we see on the outside is really just things such as the top 10 lists from CrowdTangle. That is where we are deriving our entire assessment of what manifested, what changed, how people reacted and all these things from. We have such limited visibility into what actually happened on that front. It certainly seems like some of the decisions that have been made have been made in the interests of preserving that engagement-based business model and the reversion to the status quo, but, again, we are really just going off a lot of these top 10 lists and things like that. We do not have very much visibility into what is actually happening, what people are seeing, what they are engaging with and what they go on to do next. What they go on to do next is another key piece of this. We have no visibility into that at all.

The Chair: You have mentioned several times today harmful content that originates on one platform and then migrates on to other platforms. I know that was another issue that was raised in the *Long Fuse* report. If you are creating a regulatory system, which is what we are looking to do in the UK, should you have a duty to co-operate in there, so that when something bad is happening and a company can see it, they should highlight that to the regulator or to another company and say, "Look, we, Facebook, have seen a lot of traffic coming from another site that's feeding certain Facebook groups that we think are problematic, and we're letting you know that this is going on"?

***Renée DiResta*:** My understanding is that, for certain topics, they do that. Again, it is the question of how you scope that, recognising that,

particularly for areas like political freedom of expression, people use these platforms to organise in very legitimate ways as well.

Q108 **The Chair:** Laura, I have a final question for you. You mentioned scam ads earlier. We, as a committee, are considering what the scope is for that. Again, the harm area is not just the fact that people are trying to commit fraud and mis-sell stuff to people. The nature of scam ads is that they are probably trying to mis-sell financial products to some of the most financially vulnerable people, who are more likely to believe it. They can do that only by using ad tech tools in order to reach those audiences. Do there need to be not just codes of practice but codes of practice here that the regulator should oversee?

***Laura Edelson*:** Yes, absolutely. There are two things I just want to make sure everyone is really clear on. The boundary between organic and advertising content is really porous. These kinds of scams get promoted through organic content, and then that organic content gets boosted and it technically becomes an ad, but even while it is an ad it has engagement interaction that drives further organic engagement. That is one piece of it. It can spread in a wide variety of ways. It spreads cross-platform, and it spreads through a variety of means.

Just to touch on your point about how scams are often marketed to vulnerable populations, ad tech can be really powerful in helping scammers to identify those vulnerable populations. There are a couple of things I am thinking about right away. We tracked a variety of really scammy investments that were marketed to seniors in Florida; this was reported in both 2019 and 2020, I believe. One of the ways they got on our radar is that one of the ways they were marketed was through political identity. In the United States, political identity is a really strong factor of affinity. They were marketing precious metals scams. They were encouraging people, "Cash out your retirement and invest all your money in silver, because this is the Trump investment". That was one thing.

Something that we have less information on, but it leaks into data that we have, is other kinds of scams getting marketed to groups. Veterans are unfortunately a common one. Veterans' groups get financial products marketed to them that are really scammy. Again, they are presented as some kind of government programme for veterans and they are just scammy loans. Again, some of the ad targeting tools are really effective ways of identifying these vulnerable populations. They can be long effects, whereby you may use these ad tech tools to identify this population of vulnerable people.

We saw anti-vax content being promoted to pregnant women in the United States, where they would first promote content about pregnancy, then later they would promote their anti-vax scams as a means to sell their vaccine harm reduction supplements. It is all a vicious cycle.

The Chair: You said at the beginning that a lot of adverts on sites such as Facebook start off as organic posts, which are then boosted through paid promotions. As you say, that distinction between advertising and

content that exists in other media does not really exist in the same way. If we had a regime where we regulated organic posting but not advertising, we would actually be creating an incentive for people to boost the post through advertising.

***Laura Edelson*:** Yes. I would strongly encourage you to include advertising in what is being overseen, because there is no easy distinction to be made. If you were to make that distinction, you would just be encouraging people to boost this content. It would be a really perverse incentive.

Q109 **Dean Russell MP:** We will be speaking to the platforms in the coming weeks. There are lots of topics that we need to look at, particularly safety for children. What is the key question that we must ask platforms that we have not yet had an answer to from other committees and other investigations?

***Laura Edelson*:** I would be most concerned about the degree to which platforms collect interaction data about minors. We know that minors' brain development is not complete. This is widely known. Minors have less impulse control. The interaction data gathered about minors has the potential to be really powerful, and I would just like to know much more about that.

***Renée DiResta*:** I do not work on the children issue. I do not have the best base of expertise to draw on. I would defer to others on that question.

The Chair: That concludes our session. Thank you to our witnesses, as I said at the beginning, particularly for accommodating us in the UK time zone, by getting up very early in the morning.