HOUSES OF PARLIAMENT

Joint Committee on the Draft Online Safety Bill

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

Monday 18 October 2021

4.35 pm

Watch the meeting: https://parliamentlive.tv/event/index/348d8d04-3876-4220-a40f-906f210686fe

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

Evidence Session No. 7 Heard in Public Questions 128 - 134

Witness

I: Sophie Zhang, former Facebook employee.

USE OF THE TRANSCRIPT

1. This is an uncorrected transcript of evidence taken in public and webcast on www.parliamentlive.tv.
2. Any public use of, or reference to, the contents should make clear that neither Members nor witnesses have had the opportunity to correct the record. If in doubt as to the propriety of using the transcript, please contact the Clerk of the Committee.
3. Members and witnesses are asked to send corrections to the Clerk of the Committee within 14 days of receipt.

Examination of witness

Sophie Zhang.

Q128 **The Chair:** In the final panel of today's evidence session, we are pleased to welcome Sophie Zhang, who is testifying remotely to the committee.

Sophie, as an introduction, you are a former Facebook employee and you have been a whistleblower who spoke out against practices that you saw in the company. For the benefit of the record and for people who are following this, your role at Facebook was principally as part of civic protection, looking to identify networks of inauthentic accounts that were co-ordinating and spreading information on Facebook in different countries around the world. Is that a fair description of your principal function at Facebook, or is there something you would like to add?

Sophie Zhang: I would like to add to that. Although that is what I was doing, it was in my spare time, and I was essentially moonlighting. My actual job was as a data scientist on the fake engagement team, which is focused on the prevention of inauthentic activity, but the inauthentic activity was predominantly non-civic because most people are not politicians and most discussions are not political. I hope that makes sense.

The Chair: Yes, absolutely. It would be fair to say that, from what you have written and said, the bulk of this activity within the company is directed at preventing and removing spam content rather than necessarily protecting society.

Sophie Zhang: My team was intended to prevent spam. Although our mandate was defined rather broadly, that was the intention and area of the team. I was essentially moonlighting in a separate area that was technically in my purview, but that I was not expected to do.

The Chair: When you initially spoke out about this and wrote about it on an internal blog post at Facebook that was quite widely reported, you highlighted a number of co-ordinated campaigns of inauthentic accounts that had been used, particularly in Honduras, Brazil and Uzbekistan. Those networks were being used to spread disinformation within those countries that could have had a negative effect on democratic society or influenced the outcome of elections. Is that correct?

Sophie Zhang: That is correct, with the caveat that I think you meant Azerbaijan instead of Uzbekistan. I did not do work on Uzbekistan.

The Chair: Sorry, you are quite right. I got my "stans" mixed up.

This is a particularly pertinent question for us in the UK at the moment, particularly following the awful murder of our colleague, David Amess, on Friday. Although you were looking at a system—in this case, networks of inauthentic accounts—do you believe that the way the platform works at the moment tends to radicalise opinion, to promote extremist ideas, whether through organic posting, billions of people creating inauthentic

campaigns or bot accounts, and to create divisions in societies, to undermine those societies and to allow the distribution on an unnaturally large scale of extreme ideas?

Sophie Zhang: Before I answer that question, I will try to break it down to make some distinctions, because it sounds like you are concerned about hate speech, misinformation and other ideas that have increasingly radicalised people and resulted in incidents like the extremely tragic murder of an MP. That is distinct from inauthentic activity, because extremist content, hate speech and misinformation is a function of content, as in what the person is saying. For example, if someone writes on social media, "Cats are the same species as dogs", that is misinformation regardless of who is saying it; it is not very good misinformation, but it is still misinformation. It does not matter if the Prime Minister said it or a cat fanciers' club said it.

In contrast, inauthentic activity is a function of who the person is; it does not matter what the person is saying. If I create tens of thousands of fake accounts on Facebook and use them to spread the message "Cats are adorable", that is a perfectly legitimate message except for the fact that I am using fake accounts to spread it, and ultimately Facebook would be correct to take it down regardless of how much I yell afterwards that Facebook is censoring cute cats. Those two areas are commonly confused with each other.

There exists a public perception and stereotype that fake accounts are used to spread misinformation, and that a considerable proportion of misinformation is spread by inauthentic activity. Like most stereotypes, my personal experience is that this is incorrect, and that most misinformation and hate speech is spread by real people who, tragically, genuinely believe it, and inauthentic activity—fake accounts and so on—is used mostly to spread activity that is otherwise benign in the realm of discussion.

I want also to differentiate several other types of inauthenticity that people might be confused about. A common accusation is that people spreading misinformation or hate speech do not genuinely believe it but are doing so for their own purposes. That may be true, but it is entirely separate from the social media inauthenticity that I am talking about, because, in those cases, people are still saying things under their real names. Now that I have broken down the question and spoken about it, I am actually going to answer the question.

The Chair: Great.

Sophie Zhang: Apologies for that digression. It is not controversial to say that Facebook, and social media in general, has rewritten the rules about how information is spread and distributed. In the past, when topics of discussion were to go public, there were gatekeepers. For instance, the established media would decide whether to report on it or not. If you said that the moon was made of cheese or something absurd like that, it did not matter how much people believed your claims because the

established media would not report them seriously, so it would be very difficult to get your claims out.

Today, with social media, those gatekeepers have broken down. I do not think it should be controversial. It is a fundamentally conservative idea, the idea of Chesterton's fence: not all changes are good, and when you want to make changes you should sometimes understand beforehand what the ramifications of those changes are and why the existing system is in place for a reason. The breakdowns of the gatekeepers have had positive effects as well. Certain types of speech were taboo in the past. For instance, LGBT issues were not widely discussed in the media and in public a mere 50 years ago.

At the same time, the breakdown of gatekeepers has allowed for the increased distribution and spread of radical and damaging ideas. Today, people are concerned about free speech with regards to what you can post on social media, but I see this as a smokescreen and distraction, because the concern is not free speech but rather freedom of distribution. In the past, if neo-Nazis were allowed to speak out, people were not worried that their ideas would spread widely and be disseminated and reach others, but, today, that concern exists. No one has the right to freedom of distribution. Just because the *Guardian* does not want to publish you does not mean that you are being censored.

I want to be clear that this is not the area of my expertise. Others have talked about the way social media algorithms create an incentive for people to write discussions that are sensationalist, attention drawing or emotion grabbing. One of the easiest ways to do that, sadly, is making bold claims that fall into the realm of misinformation, hate speech and the like. That was absolutely not my purview and remit, but if the committee is interested in it, I suggest considering areas to decrease virality, such as requiring social media companies to use chronological news feeds or potentially limiting the number of reshares. If someone on Facebook shares a post and you look at the shared version and share it as well, maybe after that they should have to go to the original post to share it. I hope this is making sense.

The Chair: Yes, absolutely. Where you saw cases of networks of accounts spreading disinformation, hate speech or whatever it was, how often were inauthentic accounts a factor? I appreciate that you were principally looking at inauthentic accounts rather than disinformation or hate speech as categories of content. You said earlier that you think the biggest problem is real people posting that. What do you think is the role of those inauthentic accounts in boosting content that other people have created, to create a bigger audience for them?

Sophie Zhang: Again, I want to break that down, because I did not work on hate speech or misinformation primarily. To the extent that I worked on it, it was generally because others were concerned that those messages were being spread by fake accounts. As I said, that is a bit of a stereotype. Like most stereotypes, I do not see any evidence for it to be correct.

I will give an example from the United Kingdom, so it may be familiar to you. It was a case I worked on of not hate speech but misinformation. In late 2019, in the lead-up to the general election, there was a piece of misinformation that spread widely in relation to the story of the Leeds hospital incident in which I believe a baby was put on the floor.

The misinformation went something along the lines of, "I have a good friend who is a senior nursing sister at Leeds hospital and this is correct", et cetera. It was spread around by being copy-pasted by many different people who did not all have good friends in Leeds hospital. When this came up and was quickly debunked, it was very concerning, and many people alleged that it was spread via fake user accounts. I believe that Marc Owen Jones suggested that, for instance. That was something I was put on to, initially to look for the possibility of fake accounts being used to spread it. It was something that I and others looked into, and we did not find any evidence of fake accounts.

I want to be clear that not finding evidence is not the same thing as being sure that it does not exist. In the same way that a police officer would never be able to establish for certain that someone is not a criminal, you could always argue that they are hiding extremely well and have simply hid their misdeeds long enough. I worked on many cases of hate speech or misinformation—mostly misinformation—that were alleged to have been spread via fake accounts. Essentially, in all of them, I did not find any notable fake accounts.

The Chair: I want to ask about some of the things you worked on directly. Take, for example, the network of accounts being operated in Honduras to favour the President of Honduras. You were very concerned about that. You raised it with Facebook, and it took nine months for it to be addressed.

Sophie Zhang: It took 11 and a half months. It took nine months to start the investigation.

The Chair: You said you took that up to the vice-president level within the company. Who were the most senior people you spoke to about that in your attempt to get the issue taken seriously?

Sophie Zhang: I personally briefed vice-president Guy Rosen on the issue. Guy Rosen is the vice-president of integrity at Facebook.

The Chair: After you briefed him, it would appear that it was not enough for him to take any action.

Sophie Zhang: The general trend that I would describe is that everyone agreed that the situation was terrible, but people were not convinced that it was worth being given more priority for Facebook to act. There was mostly agreement that it was terrible but no agreement on what action should be taken and how much of a priority it should be.

The Chair: They agreed it was terrible, but they did not think it was necessarily worth Facebook's time or investment to do anything about it.

Sophie Zhang: That is the way I would describe it. At least it was not to do anything about it in a timely fashion, because it was taken down, even if it returned immediately afterwards.

The Chair: Is that because not only does it involve resources to take it down but those fake contents could be driving engagement with the platform?

Sophie Zhang: I do not believe it was an area of concern, because it was a minuscule fraction compared with the overall amount of activity on Facebook. There were thousands of fake accounts, which sounds like a very large number until you realise that Facebook has something like 2 billion or 3 billion users.

The Chair: Yes.

Sophie Zhang: I doubt that the idea even crossed their minds, to be perfectly frank. With regards to the reluctance to prioritise it, my guess is that it was due primarily to the time of taking it down, but those are perhaps political considerations because it was, after all, the President of a nation, albeit a very small one.

The Chair: Given what you said before, it sounds like you had concerns about the resources Facebook had—the number of people involved in checking content. You complained about the fact that you were often making decisions on your own about what should and should not be done.

First, do you think the company needs to put more resource into this? Secondly, the *Wall Street Journal* reported yesterday that Facebook is too reliant on AI for content moderation, and that Facebook's AI systems only catch a very small single-figure percentage of the sort of harmful content that should be removed. What are your thoughts on that?

Sophie Zhang: Absolutely. Just to break it down, with regards to the use of AI, the vast majority of Facebook moderation for content-based matters is done using artificial intelligence. By that, I mean, for instance, hate speech and misinformation, as well as spam, people trying to sell you things online—often scams—nudity and pornography, and website links that send you to malware websites. Those are relatively easy to moderate with AI, but there exist differences, for instance, in the level of enforcement between nations. If you want an AI to determine whether a content is hate speech, you need an AI that can speak that language, or at least have data in that language to classify, and, of course, resources differ considerably between nations. In addition, the definition of hate speech at the company may not agree with the widely held public definition.

For instance, as of a year or two ago, according to Facebook's policies, the phrase "Men are trash" was hate speech and Holocaust denial was not hate speech, which, I would hazard a guess, very few people agree with. I did not work on hate speech, so I do not know the other factors at

play with regards to researchers' complaints that the large majority of them were not caught.

Another concern that I would express regarding hate speech, which others have also expressed, is that the company's focus on driving down the total volume of hate speech is not necessarily the way to go, in that the risk of hate speech, ultimately, is not that many people will see it but that some people who see it will very often become radicalised by it. Others have proposed focusing instead on the people who see very large amounts of hate speech and other extremist radicalising content on a day-to-day basis and focusing on that number specifically. That seems like a good idea to me. I am very sorry; there was another part to your question and I have forgotten what it was.

The Chair: That is fine. You think that is technically easy for them to do, and rather than looking at hate speech as a total thing, say, "We can identify people who are heavy consumers of hate speech and have been radicalised by it". You think that is something that they have the technical capability to do.

Sophie Zhang: They probably have the technical capability to do it, for instance, in English. It seems a bit unlikely to me that they had the technical capability to do it in every language, although they could increase it very quickly. Ultimately, it takes resources to do that. Teams that work on integrity, investigations and takedowns are chronically underresourced, which is a statement on the company's priorities. You do not hear about the ads marketing team at Facebook being chronically underresourced, for instance.

The Chair: Indeed.

Q129 **Dean Russell:** Thank you, Ms Zhang, for your testimony today. One of the parts that is core to this Bill and that we need to get right is the legislation to make sure that organisations—specifically Facebook in this instance—do the right thing. My question to you is about the culture. You mentioned that you went pretty much to the top to raise concerns about democracy. Would you say that Facebook has a culture that would rather protect itself than protect democracy and society? If so, how robust do we need to be in this Bill to make sure that it follows the rules rather than potentially create loopholes that it will work around?

Sophie Zhang: Absolutely. I would like to take a step back and remind people that we are asking whether a company whose official goal is to make money is more focused on protecting itself and its ability to make money or protecting democracy. We do not expect Philip Morris tobacco to have a division that reimburses the NHS every time someone gets lung cancer and needs to be treated. We do not expect Barings Bank to keep the world economy from crashing. That is why Britain has its own bank.

It is important to remember that Facebook is ultimately a company. Its goal is to make money. To the extent that it cares about protecting democracy, it is because people at Facebook are human and need to

sleep at night, and also because, if democracy is negatively impacted, it can create news articles that impact Facebook's ability to make money. That said, I have several suggestions about changing the culture at Facebook, or at least creating measures on the company, with regards to Ofcom regulating the company.

The first is requiring the company to apply policies consistently, which is, I believe, in Clauses 9 to 14 of the Bill. The idea that fake accounts should be taken down was written into Facebook's policies. I saw that there was a perverse effect, in that if I found fake accounts that were not directly tied to any political leader or figure, they were often easier to take down than if I found fake accounts that were. That created a perverse effect in that it creates an incentive for major political figures, essentially, to create a crime openly. If a burglar robs a bank, the police would, hopefully, arrest them very quickly, but suppose a burglar robs a bank and that burglar is a Member of Parliament who is not wearing a mask and openly shows his face, and the police decide to take a year to arrest him because they are not sure about arresting a Member of Parliament. That is essentially the analogy with Facebook.

Others have made a proposal to require companies over a certain size to separate product policy and outreach and governmental affairs, because, at Facebook, the people charged with making important decisions about what the rules are and how the rules get enforced are the same people charged with keeping good relationships with local politicians and government members, which creates a natural conflict of interest. Facebook is a private company, but so is the *Telegraph*, the *Guardian* and so on. Those organisations keep their editorial department very separate from their business department—at least I hope they do. The idea of the *Telegraph* killing a story because it made a politician look bad is unthinkable, at least to me, and I hope it would be to other members of the committee, although, of course, you know better than me.

Dean Russell: Would it focus the minds of the senior leadership in Facebook if they were liable for the harm that they do both to individuals and society from what happens within Facebook? For example, would the situation you shared earlier about the elections have happened not in 10 months but perhaps overnight if they were liable for the impact of that?

Sophie Zhang: Potentially, but it depends on precisely how they are liable and how the rules are enforced. What I mean is that the Online Safety Bill, as I understand it, is focused on liability for harm in the United Kingdom, which is an approach that can make sense for the United Kingdom as it has robust institutions and cultures, but of course, Honduras is not the United Kingdom and Azerbaijan is not the United Kingdom. They are authoritarian countries. I see it as highly unlikely that Honduras or Azerbaijan would take an approach that required Facebook to take down the inauthentic networks of their own Governments.

The other point is how it is enforced. I have read the text of the Bill. It took quite a while. My understanding is that the first way of enforcement is self-assessment by the company in regular reports under Clauses 7

and 19. This may not be reliable, and it may actually create an incentive for companies to avoid acknowledging problems internally. If you bury your head in the sand and pretend that the problem does not exist, you do not have to report as much to Ofcom. If you look for crime, you are more likely to find it, so companies will have an incentive to look for less.

With regards to enforcement, I have two separate proposals that may be difficult to apply, but I will make them nevertheless. The first is to try to independently verify the ability of each platform to catch bad activity by having Ofcom conduct better team-style penetration test operations on certain types of illegal activity. What I mean by that is this. If you want to find out how good each platform is at stopping terrorist content, you have Ofcom send experts on social media to post terrorist content in a controlled and secure manner and see what percentage of them are taken down and caught. You can then say, "Facebook took down 15%, Twitter took down 5%, and Reddit took down 13%". I am making up those numbers, of course. In that case, you could say, "All of those are terrible, but Facebook is the best. We need to focus on the companies that are less good at this". You could take the same approach with, for instance, child pornography.

The reverse could also be used. For instance, if you are worried about harassment, you could have people report benign content to see what is done to it if the content is incorrectly taken down. Ultimately, the goal is to take down the most violating content and have the least harm done to real people. You could stop everything bad overnight by banning social media in Britain, but that is obviously not what we want to do.

The second proposal that I would make is to require companies to provide data access to trusted researchers and provide funding for such researchers to have more independent verification. However, this creates some privacy risks. Aleksandr Kogan, after all, was also a university researcher.

Dean Russell: Indeed, he was. Thank you.

Q130 **Lord Knight of Weymouth:** Thank you very much, Ms Zhang, for appearing before us and, indeed, for reading the whole Bill. That is very impressive. Should the Bill be amended to include in-scope disinformation that has a societal impact as well as an individual one?

Sophie Zhang: That is a very difficult question. Right now, it would presumably fall under Clause 46, which details the banning of content harmful to adults. My concern is how you define that. These definitions are highly subjective and may be difficult for companies to determine. Right now, I think they are based on a company's definition of what it believes, which creates an obvious gap for Ofcom enforcement in that companies can argue, "Well, we don't think this is bad". I do not know the legalities involved in the regulation. I would note that, for most social media platforms at least, the use of fake accounts especially to spread inauthentic messages is already banned.

The question is more on enforcement. There are many laws that are not fully enforced. I believe it is illegal to wear arms and armour in Parliament, but presumably there are not guards at the door checking for it in this modern day and age. Part of the issue is that this committee is naturally focused on Britain. Where I found the most harm was predominantly not in Britain but in countries where authoritarian Governments were creating activity to manipulate their own citizenry. With regards to activity in Britain, this could be targeting, for instance, foreign inauthentic activity.

Ultimately, I do not know that that is the best approach. It may be a better approach, for instance, to require companies to co-ordinate closely with MI5 or MI6 in defending Britain's security if that is the specific concern. I am not a regulator and I am not a legislator. I do not have good familiarity with the issues involved in a topic as big and potentially subjective as banning disinformation.

Lord Knight of Weymouth: Thank you. The Bill, as you will recall, imposes duties to protect content of democratic importance. I am interested in how you think a company like Facebook might interpret that, particularly given that the misinformation and fake content that you have been working on, in my view, damages democracy. You could interpret the duty to say, "We should allow all political content, because that is safeguarding democratic importance", or you could say, "No, we need to work harder on fake accounts in order to protect democracy from harm". In which sort of direction do you think a company like Facebook will go?

Sophie Zhang: I think Facebook would interpret it in a way that favours what Facebook is already doing. In this context, it would turn into protecting and disseminating content that contains information on, for instance, when to vote, what the elections are and where the voting locations are, and potentially protecting controversial content by public figures and politicians with the official justification that we should allow people to speak out openly when they are important figures. That is essentially what Facebook is already doing. I hope that makes sense.

Lord Knight of Weymouth: It does. You talked right at the beginning about the difference between freedom of expression and freedom of distribution. A lot of the discussion is that the response of platforms should be to take stuff down, but clearly there are other actions that can be taken to prevent the amplification of content and prevent things being shared.

Do you have any advice for us on the sorts of things that platforms can do, short of takedown, so that they are protecting freedom of expression and political content but also protecting us from harm?

Sophie Zhang: That is a thorny question. What companies can do theoretically involves things like reducing the distribution of certain types of content by making them seen by fewer users. This has, of course, raised concerns and controversies over what people call shadow banning

in the United States. I do not know if you have heard about it in Britain, but at least in the United States it is somewhat controversial.

Ultimately, it is not always very reliable either. For instance, when misinformation gets fact-checked and then has its distribution reduced, the fact-checkers do not have the time to fact-check every single piece of content, so they naturally focus on what is popular. When something is fact-checked and shown to be misinformation, its distribution is reduced and the fact-checking label appended, "This has been fact-checked by this organisation. You can see it here". The issue is that by the time that has happened the content has already been popular enough to be fact-checked in the first place—

Lord Knight of Weymouth: Sorry to interrupt, but is it viable to require platforms to distribute the fact-check information back to where they know the erroneous content was shared, so that people can say, "I did see that, but now I see that it was false"?

Sophie Zhang: It would definitely be possible. My question is whether it would be useful. There has been research done that showed that sometimes when content is fact-checked, people do not believe the fact-check and then start digging in their heels. My concern with that approach is that it is focused on action to reduce the distribution of misinformation, but in that case the distribution is reduced when the content has already become popular oftentimes, so it is the equivalent of closing the barn door after the animal has escaped.

It is a difficult question, because, fundamentally, companies cannot adjudicate every piece of content. You probably would not want them to do so either. Ultimately, that is why my proposals and suggestions have fallen more along the lines of reducing virality in general by reducing reshares—for instance, by requiring people to go to the initial post to reshare a piece of content rather than its being reshared and resharing it again, and going to chronological news feed rankings. The problem at hand is not that the content is being made in the first place, but that it is being seen and widely distributed, and people have an incentive to make potentially sensationalist claims.

Lord Knight of Weymouth: Thank you.

Q131 **Baroness Kidron:** Hi, Sophie, and thank you for your contribution. It is absolutely fascinating. I want to go back on a couple of things you said. Right at the beginning, you gave us a fantastic explanation of the difference between hate speech, misinformation and inauthentic accounts. For the record, could you say what you think are the primary harms relating to inauthentic spread of information? Where is the harm?

Sophie Zhang: To be clear, when you say inauthentic spread of information, you are speaking about inauthentic activity, not information.

Baroness Kidron: Activity. Indeed.

Sophie Zhang: There are several types of potential harm. There are several types of inauthentic activity that I will broadly break down. The word "bot" in the modern day is used to describe two very different types of activity: literal bots, which are computer scripts that have no real human behind them; and groups of people sitting behind a desk who are paid to do something—for instance, Russian bots. These are activities that differ considerably in type, scope and behaviour. Scripts are very good at creating activity in very large volume and very bad at creating activity that is actually intelligent or smart. If the committee were to replace its staff with computer-generated reports, it would be able to generate a large number of reports that would be completely useless, which is perhaps a good analogy with the impact of scripted activity.

I have not seen any troll farms—essentially, networks of paid users—that are run out of the United Kingdom, which is not to say that they do not exist. It is possible, for instance, that they are hiding very well, because Facebook and other companies pay a lot of attention to the United Kingdom in a way that they do not to countries such as Nigeria or Honduras. At the same time, it is true that Britain has more of a culture that does not accept such activities. Furthermore, phones and labour are expensive in Britain. In India, people can buy a Jio phone for the equivalent of £10 or £15 and someone can be hired very cheaply. That would be far more expensive in the United Kingdom, of course.

Going back to the actual question, I found very few types of inauthentic activity in Britain. The main case, as I already described to the committee Chair, was that in 2019 in the lead-up to the British general election, a candidate for Parliament received a large number of fake follows from Bangladeshi fake accounts. I want to be very clear right now that this had absolutely no effect on the outcome of the election in my personal expertise and view.

With that said, what are the potential reasons for the possible effects? To me, the main concern is an increase in credibility. Britain has a multiparty electoral system. In 2019, pro-European voters needed to decide, if they did not want the Tories to win, whether to vote for Labour, Liberal Democrats or Greens. Conversely, people who were Eurosceptics needed to decide whether to vote for the Tories, the Reform Party, or the Brexit Party, or whatever they are calling themselves now.

Baroness Kidron: Are you saying that the harm is that someone appears to be more popular than they are?

Sophie Zhang: Exactly. The appearance of popularity is important in a multiparty political system like the one in Britain. If someone needs to decide whether to vote for one candidate or another who both share their views, and they want to know who has the best chance of winning in their constituency, one way they would do so, presumably, is by looking on Facebook. If someone has 1,000 followers on Facebook, that is very different from someone who has 4,000 or 8,000 followers in terms of credibility.

Baroness Kidron: Right.

Sophie Zhang: There are other aspects of inauthentic activity that we have heard about. For instance, the spread of Russian inauthentic activity to further certain narratives has received a lot of attention. My assessment is that they have had too much attention sometimes, and I will use an anecdote to illustrate that.

In the lead-up to the 2019 election, there was a case that people in this committee may be familiar with of what were called "Boris bots". Those were alleged to be fake accounts posting certain messages in support of Prime Minister Boris Johnson in response to his Facebook posts—messages such as "Brilliant Boris 100%", "I support Boris 100%", et cetera. Those were not bots; they were real Britons who believed that it would be extremely funny to troll their political opponents by pretending to be fake accounts for the purpose of arousing fears. Of course, it raised attention in Britain. I believe the BBC eventually wrote an article about it. I was asked to urgently investigate it something like six times. After the first two, I gave up because it was very clearly the same thing over and over again.

Baroness Kidron: I do not want to interrupt you, but I want to get to the question of harm, because I noticed in the notes that when you went to Facebook they said that it was a minor matter and not important. I am trying to get to why it is important rather than why it is not.

Sophie Zhang: There are different types of importance. It degrades the democratic conversation. It harms the civic discourse. If people do not know who to trust online, they are unable to trust anything at all, and that can be very harmful. In a society like the United States or the United Kingdom, we take it relatively for granted, but in authoritarian countries you do not know if people are really who they say they are. Perhaps they are informants for the Government. Perhaps they are fake people who are paid by the Government. That is presumably part of what is going on in countries such as Honduras and Azerbaijan.

I would compare it to the paid crowds of the eastern bloc of yesteryear. When Ceausescu gave his final speech in Romania, he spoke to a crowd of 100,000 people, who were mostly rounded up, bussed in and given placards to support him. That crowd turned on him in the middle of his speech and began the Romanian revolution. When you need 100,000 people in real life, you need to get actual people. There is no way for 1,000 to pretend to be 100,000 in the real world. It is extremely hard to control that number of people. On the internet, it is very easy for a small number of people to pretend to be a very large number of people.

Baroness Kidron: That is very helpful. I have one other question, which is on scale. You said earlier that Facebook was not that bothered because it was such a small amount of the overall population, but when you have 3.5 billion, suddenly a very small percentage is enormous. Are they taking seriously at Facebook what you consider an automated harm?

Sophie Zhang: I would say that how seriously they take it depends on multiple factors. Ultimately, Facebook is a company that cares to the extent that it impacts its ability to make money and people need to sleep at night. I want to draw a distinction in that most of Facebook's investigations happen in response to outside reports and claims. Perhaps MI5 goes to Facebook and says there is something odd going on in Britain. Perhaps in a small country an opposition group goes to Facebook and says, "There's a strange group. We don't know what it is. We think it's fake". Perhaps an NGO goes to Facebook and says, "Can you look into this?"

Ultimately, what happens when there is an outside report is that there is someone outside the company with no loyalty to the company who can hold the company responsible. If Facebook does not want to act, they can tell Facebook, "Well, in that case, we're going to go to the *New York Times* and tell them you don't think our country is important. What do you say to that?", and suddenly it will be an important priority at Facebook. This is an actual story.

In my case, I was going out on my own without recourse to outside reports, and I was looking for unusual, suspicious activity worldwide. What I found was mostly in the global source, which I think is a statement on the low-hanging fruit there. What I mean is that, first, Facebook pays more attention to countries such as the United States, Britain and India because of the importance of those countries to Facebook, and, secondly, those countries have more robust institutions that can find and report potential strange activity. Meanwhile, the Government of Azerbaijan are not going to report to Facebook about the activity created by their own employees. Because my loyalties were theoretically to the company, so I do not think there was pressure on Facebook in the same way.

The argument that I always used was that this was so obvious that sooner or later people would notice. In Azerbaijan, for instance, even BBC Azerbaijan was a target of the Azeri Government for harassment. I always thought it was quite odd that it never noticed and reported on it, quite frankly. Facebook has many leaks. If it got out and was reported that Facebook sat on it for a year, it would be absolutely awful for Facebook. Of course, this became a self-fulfilling prophecy, because I am speaking to you about it right now, but we did not know that at the time.

The Chair: Thank you.

Q132 **Debbie Abrahams:** Hello, Sophie. Thank you again so much for providing evidence to the committee today. My question is a little closer to home. It is in relation to fake accounts that might have been used in the 2016 and 2020 US presidential elections. I understand that that escalated in the 2020 elections. Had you worked in the run-up to the 2016 or 2020 elections to identify inauthentic accounts, and how did that change?

Sophie Zhang: I want to be very clear on two things. First of all, I was hired by Facebook in January 2018, so I did no work on the 2016 elections. Secondly, there were dedicated people who worked on the US 2020 elections. I was moonlighting in that area. I was not one of them. I did not personally work on the issue. My knowledge of it is limited to what I have read in the press.

I worked sometimes on related issues. For instance, in the United States in early spring 2020—I believe it was February or March—there was a Facebook page that received attention in the American press because it was alleged to be a Russian disinformation operation. The page was spreading misinformation and, notably, it sometimes responded to critics in Russian Cyrillic. I was one of many people who investigated that, and we quickly found it to be a North Carolinian who believed it would be very funny to pretend to be a Russian to arouse the fears of his political opponents, which I suppose is something that Britain and America, sadly, have in common.

Debbie Abrahams: That is very helpful indeed. Apologies for getting my dates wrong. I am sorry if this is little bit naive as I am a non-tech person. You have expressed very clearly the difference between authentic accounts, which may not be about presenting misinformation but are fake inasmuch as they are distributed and amplified. Once those authentic accounts are distributed, do they then morph into accounts that might also provide disinformation and fake news? You have a hook into different people who might have accessed the original account. Do they change and then provide misinformation?

Sophie Zhang: I want to be very clear that this is not an area that I worked on. That said, the concept that you describe exists and I will give examples from memory. For instance, suppose there is a page on Facebook called "I love cats", and it spreads cute pictures of cats and people follow the page because cats are adorable, but suddenly one day the page changes to "I love the Tories"—sorry, Tories—or "I love the Lib Dems" or "I love Labour", and it posts content about how these are great. In that case, there is no misinformation but it is still inauthentic in the sense that the page was pretending to be something in order to gain an audience and then completely changing its message to spread it to a new audience.

Debbie Abrahams: Right.

Sophie Zhang: It is not something I worked on personally. There were other people at Facebook who worked on it. The concept certainly exists, if that makes sense.

Debbie Abrahams: Lovely. Thank you so much, Sophie.

Q133 **The Chair:** Thank you. Sophie. With regard to those last questions and the questions from Baroness Kidron, you have worked on networks of fake accounts in countries such as Brazil and Honduras, as we have spoken about. Do you think in those countries, particularly where there is

much less supervision of what goes on in social media, that Facebook could be regarded as a force that is being used to undermine democracy, inasmuch as democracy exists in those countries?

Sophie Zhang: It is ultimately a difficult question. Is Facebook being used as a tool by authoritarian Governments in those countries? Yes, it is. Is Facebook used by the opposition in those countries to get their voices out? Yes, it also is. When I came forward with the network in Azerbaijan that was focused by the Azeri Government entirely on harassing the Azeri opposition, I was a bit surprised by the official response from the Azeri opposition leader, Ali Karimli. He would have had every right to criticise Facebook and Mark Zuckerberg, and denounce them for enabling that authoritarian activity, but he did not. Instead, he said something like this—I am paraphrasing from memory: "I thank Mark and Facebook for building this platform. Facebook allows the opposition to get our voices out. With that said, Facebook should hire someone who speaks Azeri".

I am sure it would have been very tempting for him to denounce Facebook, but Facebook is important in a country like Azerbaijan, which is essentially a one-party dictatorship that is so democratic that in 2013 they accidentally released election results the day before the actual election. I wish I was joking. This is a country where the opposition does not have other significant tools, and, for all of Facebook's flaws, Facebook is still valuable to them.

Take Myanmar, for example. In Myanmar, Facebook has absolutely been used to further hate speech and has allegedly created conditions for a genocide. At the same time, it is also true that social media has been used by the people of Myanmar to co-ordinate against the latest military coup d'état this year in a way that they were not able to do for the coup d'état 20 years ago. The ultimate question about the net impact of Facebook on democracy in these societies is very difficult to answer. I hope that makes sense and does not come off as a dodge.

The Chair: No, it does not. You have been very clear as well that Facebook does not put anything like the resources it should into dealing with clearly problematic and harmful areas of content and, alongside that content, the networks of inauthentic accounts that are engaged in boosting or promoting it. On top of that, it would seem that executives in the company sought to dissuade you from investigating these issues.

Sophie Zhang: I was never directly told no until the end when I actually was told no. Most of the time, I was never told no. I would hazard a guess that it was a situation in which people did not want to have an official answer on the record that would make them look bad.

The Chair: In the statement you posted when you left the company, you said, "I was told to stop my civic work and focus on my road map on pain of being fired".

Sophie Zhang: Yes, that is what I meant when I said I was eventually told no. It was at the end of 2019 and the start of 2020. Before then I

was never officially told no, including by the vice-president. I am trying to be clear about this. I hope it is.

Q134 **Lord Clement-Jones:** Thank you very much for a fascinating session, Sophie. Do you think there is a role for a regulator in being able to insist on preventing virality? You talked about distribution virality. The kind of thing I am thinking about is a circuit breaker. Should the regulator have the power to insist on that, or is it a tool that should be expected of a platform?

Sophie Zhang: My initial reaction is somewhat leery just because this could set an unfortunate precedent that could be used by other authoritarian countries. Facebook has circuit-breaker tools in countries that face threats of imminent violence. It has tools to tone down virality. But you could also imagine, for instance, a case in the Russian Federation where Russians protest en masse using social media to co-ordinate, and the Russian Government insist that social media tone down virality and inhibit activities. It is a question that the members of the committee should consider and discuss, but my initial reaction is that I am leery of setting an unfortunate precedent. The legislation contains grounds for criminal penalties for failure to comply, including prison. I am leery of that clause, because that tool has so far been primarily used by authoritarian countries such as Russia to enforce compliance.

Lord Clement-Jones: Thank you very much indeed.

The Chair: Thank you, Sophie. We are extremely grateful to you for giving evidence to us this afternoon. We will end the session now as Members wish to attend the memorial service for our former colleague, David Amess, which is taking place at six o'clock, but we are very grateful for your time and very candid answers this afternoon.

Sophie Zhang: Absolutely. Thank you very much. It was a pleasure and an honour.

The Chair: Thank you.