**NOTE OF MEETING WITH TWITTER**

**Date: Wednesday 25 October 2017**

**Time: 13:30 - 14:15**

**Location: Lord Bew’s Office, 1 Horse Guards Road**

**Attending**

Nick Pickles - UK Public Policy Manager, Twitter

Jane Ramsey - Member, *Committee on Standards in Public Life*

Stuart Ramsay - Secretariat, *Committee on Standards in Public Life*

Dee Goddard - Secretariat, *Committee on Standards in Public Life*

JR: Welcome. Lord Bew is unfortunately unable to attend today. What is the responsibility of social media platforms for online intimidation and abuse?

NP: We see the value of free expression, but people also need to be able to express themselves free from being shouted down. We want candidates during elections to express themselves freely and safely. Twitter is used by journalists and politicians, and we don’t want them driven off social media. It’s a very high priority for us. Our CEO has said that safety is a number one priority. In the last two weeks, we have added considerably to the list of improvements we have made. We realise we need to do more to enforce the rules, and make sure people feel safe on the platforms. We’re making rules clearer, strengthening them in some cases, enforcing them better. We’re dealing with breaches through automation. The election brought to bear the challenges around free expression and abuse that we’ve got to deal with.

JR: What are the improvements?

NP: If we change the rules, people change their behaviour to get around it. For example, we used to prohibit direct violent threats, and people then sent indirect threats. We’re strengthening the rules on uninvited sexual advances, on non-consensual nudity - such as upskirt shots, as these are used to harass people. And we’re strengthening enforcement. I can give you a full write up of the calendar. You can read the internal conversation too, actually, as it’s been leaked. But we also are thinking about how we don’t over-correct too much - and reviewing that process now. Following two or three big announcements - one of which is how we nudge people back to good behaviour, people who are in a ‘penalty box’ who can’t tweet for 12 hours. We’re seeing good progress on people who go through that process who don’t go back to it again. I can send you more detail.

JR: That would be helpful. What’s your position on anonymity?

NP: It’s more complex than whether you’re anonymous or not. An important part of our platform is not being able to use your real name - parody accounts, or accounts in support of someone, such as @JC4PM. Or where using your real name might threaten your safety - whistleblowers, or human rights activists in the Middle East. Second, the vast majority of people have given us information about who they are - like phone numbers. 80% of UK users access Twitter on their phone. If you give the police a phone number they can easily identify someone. Something we’re doing aggressively is pushing people, saying ‘if you don’t give us your phone number, you can’t come back on the platform’. It reminds people they’re not anonymous. But Korea introduced a rule banning anonymity - but it made no difference to abuse and was abandoned.

JR: Do you have plans to make more people give you their phone number?

NP: If they breach the rules, or the technology detects that there is abuse. For example, if A blocks B and then B sends A ten or twenty tweets in a short period. In that case, we might restrict their tweets only to their followers. We’re using technology in ways to try to find that behaviour. User reports are still critical - and we won’t get past that because context is everything. We think there’s good progress. If they don’t give us their phone number and break the rules, we’ll kick them off.

JR: Would you restrict accounts only to those who give you a phone number?

NP: No. Not everyone has a phone, rural areas have poor mobile phone signal, and there needs to be protection for human rights activists. We need to balance a right to pseudonymity with safety values, and we’re using technology in different ways to strike that balance.

JR: How do you respond to the criticism that platforms like Twitter allow a huge amount of abuse to be made directly and anonymously to individuals?

NP: The overwhelming majority of people do not use Twitter to abuse people. The General Election was incredibly polarised. For Brexit or against, for Jeremy Corbyn or against. That polarisation led to people being abused. But the question is where the line is between abuse and political debate. There’s been a lot of reporting around Labour MPs who are not supportive or Jeremy Corbyn. But when does political debate in the Labour Party become abuse? We can’t say where the line is. Different people will have different thresholds. But the idea we don’t have rules or don’t enforce them is nonsense.

JR: Would you increase monitoring in election periods in order to maintain a healthy democratic culture?

NP: We want people to express themselves for a whole Parliament, not an election. This election did seem more polarised, more aggressive. Whether that was the print media, people defacing posters with swastikas - which I thought was remarkable in this day and age. But people have a right to express themselves and they shouldn’t be silenced through abuse. The solutions that are required that go far beyond social media.

JR: Could you talk us through the process of how you monitor and deal with content that breaks Twitter’s rules?

NP: It was twelve clicks to report a tweet, it’s now three or four. We’re using technology to identify things without user reports. The headline figure is that we’re taking action on ten times more accounts than this time last year, due to internal machine learning. We have a dedicated reporting flow for violent threats, and for hateful conduct based on the UN Declaration protected characteristics. We will use technology to prioritise the reports - for example, something that is two weeks old, that is reported by someone who’s not mentioned, is lower priority. It is automatically triaged. This is the most complex challenge. But we are also essentially adding warning labels to offensive messages, where we think it is offensive but won’t break the rules. It can hide offensive replies. It will have false positives but it is there. We’re working on changes to make clearer in emails and in-app exactly which rule was broken, on which account. Currently users don’t get told what rule they broke. If you’re warned, we walk you through a process where you have to delete the Tweet. It’s been likened to a ‘speed awareness course’, to make sure people have understood the rules - it’ll be on the screen, in the text.

JR: Is it three strikes and you’re out?

NP: It depends upon the infringement. A violent threat, we might suspend on the second report. But if two people are arguing and it gets out of hand we might be more lenient. But we might still suspend someone for three separate breaches.

JR: What about content that is replicated through screenshots?

NP: People will often screenshot behaviour to condemn it. Often we will take it down but the screenshot will stay on much longer. I think it is important that people call out behaviour - but yes, the platform will keep content that breaks the rules through the screenshot.

JR: it would be helpful to quote an example of that - could you give us an example?

NP: There are many out there.

JR: About elections?

NP: I see this quite a lot - including something like an ISIS tweet. We might take it down if it has private data, like a phone number or an address. But the screenshot and content which breaks our rules is not the same...

JR: Because of the intent?

NP: Exactly. If you look at Stella Creasy and Caroline Criado-Perez, or Everyday Sexism - calling out behaviour is part of changing behaviour. If you are praising it, we might take a different view. I did see one tweet where a screenshot was reported, but the report was about that content, not the tweet - which was critical. We want people to be critical of content they don’t agree with, as that’s what drives change.

JR: It isn’t our job to tell you what resources you deploy, but what assurances you can give about whether you deploy significant enough resources to identify, assess and remove content where it breaches your own policies? You clearly collect data - like the ten times more accounts taken action against - what transparency is there over the number of complaints, the timeframe for triage, and how you’re monitoring your timescales on takedown?

NP: We need to be more transparent on this. I was asked this question this morning on counter-terrorism. We’ve removed 935,000 accounts on terrorism worldwide - but we were asked ‘it’s a big number, but what does it mean?’. The number of reports is a very crude number. We saw a mass reporting campaign about a picture of a cat. We definitely see reports that are intended to silence people that they disagree with - political activists, campaigners, they will definitely try to report their opponents. It’s like saying ‘how many crimes were reported on 999’ - you can’t sort out how many were unique reports and how many are cats stuck up trees.

JR: The ambulance service do count reports, and how many led to paramedics coming out, or the air ambulance, or how many go to A&E, so you can follow patients through. They are hard-won statistics to achieve, but they are visible.

NP: That is interesting. We’re working out what are the numbers that are useful, that have context. We have trends around events - during an election you will see more politicians reported than after an election. We’ll see spikes in reports due to press coverage. The numbers we have put out is the number of accounts we are actioning, how many are warned who aren’t warned again, and what more we can do - the users are more concerned about the personal picture, not the aggregate picture. They want personal transparency - to know what breaks the rules.

JR: They want ‘you said, we did’

NP: They want the ‘why’.

JR: And you’re making improvements on that?

NP: Yes.

JR: What about timescales - what if a candidate had their child’s address put up? How quickly would it be taken down?

NP: Crudely, as soon as possible. If I was a party, I would want it to be down within minutes. We would work with political parties - they can file high profile requests, they have my email, they have my phone number. It would be down within the hour. We can escalate these internally. During election campaigns our relationship with political parties are extremely important.

JR: We would have people say things they reported haven’t been taken down quickly...

NP: It may not breach the rule - it can be subjective. The run up to the election was almost worse for some MPs, because those who disagreed with their party’s leadership, the pre-election period there was talk of deselection and about loyalty. But then the election focussed minds and that calmed down. The quality filter is on by default - we’ll now allow you to control what you want to see, if someone hasn’t given a phone number, or hasn’t changed their profile picture. Take tweets about selling arms to Saudi Arabia - these are very intense policy debates - if you want to be part of it, fine, but if you don’t, then it doesn’t have to be part of your timeline. Technology is never going to catch everything. User reports are still incredibly valuable. I spoke to most parties most days of the campaign.

JR: Let’s stick with metrics. Say you do have something that falls foul of your rules. The NHS have targets of seeing 95% of patients in A&E in 4 hours, and so on. These are very easy targets. How each provider meets this target is published monthly. Do Twitter do that? Once it’s triaged, and you know it’s urgent, is it written down, is it available to users?

NP: That is a key question for us, how are we transparent in that way - putting something out that’s useful. The total number of global reports is not useful.

JR: I’m not talking about how many people see the doctor...

NP: That’s the conversation we’re having right now. I can send you the CEO tweets - it’s something we’re figuring out. Doing it without showing you individual tweets is hard. Not everyone will agree with our prioritisation. We’re also not at the end of what technology can do. We’re also thinking about how we can be transparent about action we take automatically, without reports. But there is definitely renewed emphasis about how we can get more transparent.

JR: It’s good to hear that.

NP: Another aspect is that if you’re not the individual, you don’t hear the outcome. We’ve started notifying in-app rather than in email. But we want to get better at letting people know.

JR: The question of your users will be, are you applying sufficient resources? But you can’t answer unless you have the metrics.

NP: Even if you publish a figure and a target, users will have a different experience. It might be that people disagree about whether something is private information - for example, someone’s address on a ballot paper: is that private information? Transparency and time and resources are important, but the policy question is an important one as well.

DG: Are you doing anything in the AI space on dogpiling or Twitterstorms?

NP: You can search on Twitter for ‘increased spike in notifications’ - you can see people tweeting in real time. If it picks up a high volume, there’s a screen - it’s fullscreen - and it asks if you want help. It’s automated but you can filter it out. What is harder is if you have a tweet that goes viral - that is disruptive. But it doesn’t measure intent - it’s automated.

DG: What about reporting?

NP: If you have something that goes viral, but it’s within our rules, when you report, you can report five tweets from the same user. But multiple different people are sometimes targeting an individual at scale. This is where they need help - and that’s why we have a relationship with political parties.

DG: Should it be dealt with directly by you or through the political parties?

NP: It depends - candidates have different views. It’s a busy time and you don’t have time to read all your emails. We’re working with new MPs, we have a short-form brochure we’re giving them. I’m meeting the Chief Whip, and we’ve got a meeting coming up with the Deputy Speaker. We have a good relationship with the Electoral Commission, though that’s not the right channel for a candidate. Candidates who are MPs can change their Twitter handle so they don’t say ‘MP’ - we work with parties to change it so they don’t lose their verified status. But there was always someone who wasn’t on the list who didn’t know. THe question is what’s the central resource you go to that is apolitical - what’s the single point of contact, where’s the guidance. It’s all in different emails, it’s not public - but something like that, a registration process where you get access to candidate resources that support you, that’s a way of making sure things don’t get lost.

JR: We’ve heard evidence of different experience for those who are standing again, and those who are new. But is this just talk, or is a plan emerging to do something about this?

NP: We’re asking, what do we need, what would it look like, where would we put it. We can’t host it - but it’s a standing offer I make to everyone I speak to, that we want to be part of it.

DG: Are you happy to work with other social media platforms?

NP: Absolutely, we’re all members of UKCCIS. We all speak together and talk about these issues regularly. But from personal experience, there’s a difference between candidates in target and non-target seats in terms of resourcing. It might be the Association is the entirety of your campaign team. But you’ll have more contacts and high-level visits if you are in a target seat.

DG: Our evidence suggests that users who experience illegal online intimidation via the social media networks frequently do not report it to the police. What steps do you take to ensure that cases of illegal behaviour online that are reported to you can be easily referred onto the police or relevant authorities?

NP: Take the Malicious Communications Act and Section 127 of the Communications Act - they are evidentially more direct than stalking and harassment. I was concerned that in context of hate crime and violence against women and girls, the two offences were being used because they were easier to prosecute. We worked very closely with the CPS on their new guidance. I think it’s absolutely right. Some candidates do have stalkers in their constituency - but we should be asking ‘is there a wider threat’ not ‘can we prosecute them for an offensive tweet’. We were speaking to them in Dublin last week, about how can you write a restraining order for Twitter, and spoke to the police today. The Mayor of London’s MOPAC - we’re part of it with the Met Police, StopHateUK, Google, and Facebook - recognise that lots of behavior isn’t illegal. But the online hate crime hub will mean that non-criminal behaviour reported can still come to us and we can deal with it if it breaks our rules. But lots of reports both aren’t criminal and don’t break our rules. If there’s a threat to a MP, they get full cooperation from us - if things go awry, it’s my phone that rings, and we fix it.

DG: What about from the other side, what information, as a user, do I get about illegality and how I report it?

NP: If you get a violent threat, the last stage of reporting is ‘do you want to report this to the police’, and you get a single email with the information that you can take to the police, and a link to our reporting centre. That seems to be well-received. If there are innovations in other areas, that’s definitely something we’re interested in exploring. Most of our rules go far beyond the criminal law. We take down a lot more than the law requires. There’s also a challenge of a single offender with multiple victims - you can have multiple forces working on the same individual. I’ve said repeatedly publicly that the APPG on Antisemitism suggested using ASBOs on low-level behaviour which is essentially online anti-social behaviour. We have the tools already there and it seems to be common sense to use those same tools.

DG: How do you respond to the view that you should take more responsibility for the content posted and drop the idea that you are simply platforms and channels for publishers? How do you deal with that discussion?

NP: The problem is that the model is publisher/platform - it is more complex. We have Twitter moments - and we’re responsible for that content. At the same time if someone creates an account and breaks our rules, we will enforce our rules but we cannot pre-moderate. We have 500 million tweets a day. The only people who make those arguments are people who would prefer if social media didn’t exist. On publishers and platforms - there is complexity here. Are there rules, yes, do we enforce them yes, should we work harder, we’re doing that. Is there a simple solution that if we are publishers all this would go away? But if we were publishers - the Canary ran a story saying Laura Kuenssberg spoke at Tory Party Conference. Are we supposed to decide that the Canary can’t have a Twitter account because of their Laura Kuenssberg story? Tech companies should not be making those decisions. The policy debate hasn’t happened yet, we’re still in the early stages of the consumer internet - people think that we are at the end. There’s a lot more complexity in the policy debate than saying, ‘are you a publisher or a platform’.

JR: Some people would say, ‘you would say that, wouldn’t you?’.

NP: Well, yes - but it depends upon your policy objective. A lot of people who say we should be publishers have a commercial interest in that as well. So you have two sides of an argument who reflect their commercial position and their view of the world. Something we do forget are users. The vast majority of users who have an overwhelmingly positive experience, are they willing to wait for their posts to be moderated? If we suggested that for candidates we’d immediately be accused of bias if some went up faster than others. It’s great strength of democratic debate that candidates can speak to the electorate directly. Leaders, world leaders, can speak to the public directly. One of the critical things that relies on is that the communication happens in step with the real world. So whether that is tragedies and national disasters, and citizens can speak to their leaders in real time, to say ‘events are happening in the real world, and we can’t wait for tomorrow’s papers for debate’. That’s a very powerful tool for democracy, and the idea that we should roll it back and go back to a world where we treat everything as published...is not a solution to some of the civic issues we have around politics, or safety and abuse. And I say that seeing the tweets about controversial comments. There is the complexity of secondary publishing - for example, YouTube keeping up a video of Fox News playing an ISIS video. A news organisation had already made the decision to publish it. There are immense risks with technology companies making decisions in that context.

JR: I obviously don’t disagree with your point about the right to freedom of expression. But one of the things you said is that Twitter is human behaviour, and you can’t modify human behaviour. But whilst recognising your points about human rights and whistleblowers, the vast majority of people when they make comments to newspapers, and angry people on television programmes - these are behaviours we would defend even if we disagree. But the difference is anonymity. I have three Twitter accounts and I can be recognised on all of them. I wouldn’t be disrespectful or rude - I hope you’d defend my right to do so - but at any rate people would know it was me. I’d posit most people who tweet about elections don’t need to be anonymous. What we need is where a candidate is getting stuff which is not illegal but falls outside your guidelines - they should get protection from that. Your point falls down because you defend anonymity.

NP: We want to change behaviour -

JR: I thought you were saying you didn’t want to modify human behaviour. In your world, it’s anonymous.

NP: Lots of people are identifiable.

JR: Do you have stats on that?

NP: This was just from things that have been escalated. In politics, people do want to have their voice heard, doing it with their own name. We have one switch where you can never see tweets for people haven’t changed their profile picture, or who hasn’t registered their phone number. It’s not there any more -

DG: It is still there though, other people can view it.

NP: But you would have to go to my profile to see my tweets. It is not public - it is not like a forum or a newspaper. On Twitter you see things sent to you, or that you follow -

DG: Or a hashtag.

NP: If you click on a hashtag. There are positive benefits to that in many ways. We want to drive people to be identifiable with phone numbers. We use that to remind people that they are not fully anonymous. You can buy a new phone, but it’s friction. We’re putting that friction in place for certain accounts - and we’re using technology to do that. The Korean law didn’t work. There’s a danger of thinking there is a silver bullet. You can fixate on anonymity - but my worry is, if you changed this overnight would the issues still be there? Yes they would.

JR: The point I was making was that you said Twitter reflects human behaviour, but I was saying that Twitter makes some behaviour easier. The evidence we’ve received is that there is something about social media permitting a behaviour, or enhancing particular aspects, for good but also for ill. And anonymity is doing that.

NP: People in government would say the same about off the record quotes. Someone using anonymous accounts shouldn’t be able to intimidate someone. But we don’t think that there are two groups, anonymous and named users. There are good actors and bad actors in both groups and we are using technology to tackle it.

JR: We had an anonymous submission - she gets horrific tweets, and she doesn’t look at her tweets, but her husband does. She lives near students and looks across and thinks, ‘I wonder if anyone there tweeted those’. She lives in fear and Twitter has enabled that.

NP: If those broke our rules and were reported, I hope they would have been taken down. These don’t come out of nowhere - candidates said they received emails with this sort of content. The same with bricks and letters. If it’s an issue of us and the police then I want to fix it. I don’t know if it’s been reported to us.

JR: It’s an ongoing case.

NP: I can make sure that the police get the assistance they need. As a company we want to understand why they feel afraid, take the person off the platform, and make sure they are held to justice in the real world. Taking them off social media does not solve the problem. This review has made people realise that there are a number of ways that people interact with Parliamentary candidates.

JR: My final question is, would you be receptive to the idea of coming together to discuss with government and political parties a way forward to try to find ways to combat and prevent online abuse and intimidation?

NP: We are. Some of the stuff the CPS has done on stalking and harassment, the new guidance, the conversations I’ve had with the Parliamentary police. Everyone has a vague idea of what needs to happen but no one has pulled it together. As a former candidate it’s something I want to see happen. I want to make sure that future candidates can communicate in the world in the way that I have.

JR: Thank you very much, Nick.

CSPL Secretariat

October 2017