Julian Knight MP
Chair, Digital, Culture, Media and Sport Committee
House of Commons
London SW1A 0AA

14 May 2020

Dear Mr Knight,

Thank you for inviting Facebook to give evidence at the recent meeting of the Sub-Committee on Online Harms and Disinformation, and for your letters following up from that appearance.

Before addressing your specific questions, I wanted to share up front some further information to assist the Committee with its inquiry, including updated global figures on our efforts to tackle coronavirus-related misinformation.

- Facebook works with over **60 fact-checking organisations** in more than **50 languages** to help us tackle misinformation. In the UK, our partners are Full Fact, FactCheckNI, and as of March, Reuters UK. Once a piece of content is rated false by fact-checkers, we show it lower in Feeds so fewer people see it, we notify people who shared it, and we cover it with a warning label that gives people more context.

- During the month of April, we displayed warning labels on around **50 million** pieces of content related to COVID-19 on Facebook, based on around **7,500 articles** by our independent fact-checking partners. The equivalent figure for March was that we displayed warning labels on around **40 million** pieces of content. When people saw those warning labels, **95%** of the time they did not click to view the original content.

- Some misinformation can contribute to the risk of imminent violence or physical harm. We work with trusted partners, including health authorities, to determine this type of misinformation, and we remove it from our platform. We have removed **hundreds of thousands** of pieces of misinformation globally in these cases.

- Since 1 March, we've removed more than **2.5 million** pieces of organic content for the sale of masks, hand sanitizers, surface disinfecting wipes and COVID-19 test kits.

- To date, we've directed over **2 billion people** to resources from health authorities including the NHS and GOV.uk through our COVID-19 Information Centre and pop-ups on Facebook and Instagram.

Address: 10 Brock Street
Regent Place
London, England
NW13FG

Throughout the coronavirus epidemic, we have prioritised sharing data with our academic partners and with Government institutions to help inform the public health response to the virus.

Through our Data For Good programme we have given our research partners at the London School of Hygiene and Tropical Medicine, the University of Southampton, and the Oxford University Big Data Institute access to privacy-protected population movement data. Using this, they have generated insights into the potential spread of the virus in the UK to share with the Government. We set out the different ways this data is protected and used on our website: https://dataforgood.fb.com/tools/disease-prevention-maps/.

In addition to this, we created dedicated dashboards using our CrowdTangle content discovery service, so that the Government teams that are working on the Government's response to COVID-19 can get real-time information about what topics and subjects are trending across the public surfaces of Facebook and Instagram.

We also provide data about content on our platform which violates our rules and the actions taken against it in our quarterly Community Standards Enforcement Reports. The most recent report, published earlier this week, is available at https://transparency.facebook.com/.

The Committee asked a number of specific questions in your letters. I would like to address these in turn.

1. What proportion of people who have seen (not just engaged with) content flagged by users or third party fact-checkers as misinformation or disinformation on Facebook and Instagram are then alerted to this and provided with authoritative information? What is the total number of users that have received this alert since it was rolled out?

When one of our fact-checkers rates a piece of content as false, we show it lower down in users' Feeds so that fewer people see it. Where it does appear, we cover the content with a warning screen that links to a debunking article by the fact-checker, which means 100% of those who see content already flagged as false by our fact-checkers will be given this additional context. As set out above, during the month of April, we displayed warning labels on about 50 million pieces of content related to COVID-19 on Facebook, based on around 7,500 articles. When people saw those warning labels, 95% of the time they did not click to view the original content.

Address: 10 Brock Street
Regent Place
London, England
NW13FG

As you note, we also send similar information to users who had previously *shared* the content that has been debunked, via a notification. We focus on these individuals, rather than those who simply may have seen or scrolled past the content, because we need to balance the importance of showing accurate information to people who may have engaged with misinformation, against drawing attention to these false narratives among people who may not have noticed them. For that reason, we do not have figures for what proportion the number of users that we notify represents out of all those who may have simply viewed the content.

Additionally, we are now showing new messages to users who reacted to, shared or commented on content we subsequently *removed* for being harmful COVID related misinformation. As the information they saw no longer exists, we can't show them a fact-checked article in this instance—instead we direct them to the WHO's mythbusters page.

We believe this balance between removing content that could cause harm, reducing the visibility of content rated as false, and showing messages to those who have interacted with any misleading content is the best way to help people get context, especially on hoaxes they may also see off our platforms.

2. Does Facebook today have the same number of full-time equivalent staff working directly in content moderation across all platforms as before social distancing measures were put in place (i.e. 1 February 2020), both for English-language users and globally?

No. Around 35,000 people work on safety and security at Facebook, and around half of those directly review content. Due to the unprecedented COVID-19 situation, we took the decision in March to temporarily send these content reviewers home, for their health and safety. As a result, since Mid-March we have been operating with a reduced content review workforce.

We announced this decision at the time alongside several changes we have made to help keep our platform safe. This included taking steps to enable some of these content reviewers to work from home, increasing the use of proactive detection technologies, shifting the most sensitive content review work to full-time Facebook employees and ensuring we continue to prioritise the most harmful content for human review.

We have been steadily increasing the number of our content reviewers who are able to work from home and we have now enabled the majority of our content reviewers to do so. Additionally, we recently announced that we have started to bring a small

Address: 10 Brock Street
Regent Place
London, England
NW13FG

number of content reviewers back to offices to help review content related to real-world harm.

3. How are you balancing moderation of different kinds of harmful material? Are there some categories of harm that are being expedited, and if so, what categories?

We have taken a number of steps to help keep our platform safe, as set out above. While we have continued to enforce our policies, we have been prioritising for review content that has the greatest potential to harm our community. This includes content on Facebook Live; content related to real-world harm such child safety, suicide and self-injury, and terrorist content; and harmful content related to COVID-19 (including content that makes false claims about cures, treatments, the availability of essential services or the location and severity of the outbreak).

4. In a letter to the Committee, the Secretary of State said that the Government has been working closely with social media platforms "to tackle misinformation and disinformation together". Can you provide more detail on exactly how Government and your company are working together, and provide information on whether there have been discussions about future plans to tackle online harms and disinformation?

Given Facebook's international presence we have been working to support Governments and international health authorities around the world since the start of the year. Facebook maintains a regular dialogue with many Government departments, including not only DCMS but also Home Office, Cabinet Office, the Department for Business, and the Department for Health and Social Care. We have been speaking to DHSC about the Coronavirus since early February, and in addition to working to promote accurate information from the NHS and the UK Government, this has included working to tackle the spread of misinformation from the start.

We hold regular meetings with DCMS and the Cabinet Office on misinformation to share updates on our work and high-level trends, including those from our third-party fact checkers. Core Government teams including the Countering Online Manipulation Unit and the Rapid Response Unit are able to report misinformation directly to us, and we meet regularly to ensure that we are maintaining best practice in handling reports of concerning content.

As the Secretary of State said to the Committee last month, one of the most effective ways to counter the spread of misinformation is to connect people with accurate, authoritative facts. This has been a focus of our efforts with Government, which is

Address: 10 Brock Street
Regent Place
London, England
NW13FG

why we partnered with Public Health England to launch an NHS Coronavirus WhatsApp bot, which has sent more than one million messages providing the latest statistics, guidance, and mythbusting resources.

By featuring links to guidance about coronavirus from the NHS and GOV.uk (as well as international health authorities) in people's search results, at the top of News Feeds, and in our Coronavirus Information Centre, we have directed more than 2 billion people globally to trusted resources, with more than 350 million people clicking through to learn more. In the UK NHS Digital has confirmed that around half a million people visited the NHS websites from Facebook in February this year, compared to 170,000 from all social media sites in January.

5. Assuming the proposed duty of care set out in the Online Harms White Paper and initial consultation response has public and civil society buy-in, would Facebook support it in principle?

Facebook has said for a long time that we feel there should be more Government regulation in the area of online content. We have contributed to the Government's thinking as they developed their online harms proposals over the last three years, and we welcomed the Government's Online Harms White Paper and the regulatory framework it proposes.

As I said during the session, the term 'duty of care' can have a number of very particular legal meanings, and the Government itself is continuing to consult on the framework, what specifically a duty of care will entail and how the duty of care will apply. We expect that the Government will set out its final proposals shortly, including details on the duty of care model, and I will be happy to provide the Committee with Facebook's position on the proposals at that point. Until then however, we look forward to seeing more detail from the Government.

6. Does your company endorse the recommendations set out by Full Fact in their report on the Third Party Fact Checking Programme, such as more data for fact checkers about the spread and flagging of content, a more in-depth rating system and an expansion of the Programme to include Instagram? Or not, why not, and which recommendations specifically?

We value the constructive feedback that Full Fact and our other fact-checking partners provide. Full Fact's 2019 report made a number of recommendations across the three themes that you refer to above, and we have made progress in all three of these.

Address: 10 Brock Street
Regent Place
London, England
NW13FG

On additional data, we send fact-checkers reports that include customized statistics that reflect the work and impact of each fact-checker. These reports were the result of feedback that we heard from our partners that they'd like more data on the impact of their efforts.

For example, we share the number of fact-check articles they've submitted, the number of pieces of content they've rated, the notifications that have been sent to users and Pages as the result of their work and examples of cases in which we've been able to identify and reduce the spread of identical pieces of content after receiving a fact-check from them.

We're also always considering how the rating options we provide could be more relevant. Any changes that we make need to reflect the feedback from our network of more than 60 fact-checkers across the globe. This is a topic of constant discussion and feedback from our fact-checkers, including when they attend our annual summits hosted at our headquarters, and any changes we make to the rating system will be informed by Full Fact's recommendations.

Lastly, on product developments, we announced the global expansion of our fact-checking program to Instagram in December 2019. However, we continue to work with our fact-checking partners on new ways to support their work across our family of apps. In the case of Full Fact, this has included discussions that are still ongoing about the use of our WhatsApp business app and API. We have also partnered with Full Fact in the media literacy space—during the 2019 election, we supported a media literacy campaign run by Full Fact on our platform to help people understand how to spot false information, which reached 11 million people.

7. How have you been working with partners in the Trusted News Initiative to ensure that authoritative information is surfaced appropriately and misinformation is demoted on your platforms? Do you adjust your systems to reflect these insights and how do the algorithms and systems work to implement what partners are telling you?

The Trusted News Initiative was set up by the BBC in 2019 and we are a member, along with other tech companies and publishers. We initially piloted a misinformation alert scheme in late 2019 to cover the UK election, whereby through an email channel publishers or platforms could flag to the whole group any content which could potentially represent the most serious forms of misinformation.

It was decided earlier this year by members to bring back the channel to combat the most serious forms of potential COVID-19 misinformation which could lead to real world harm, and any false accounts purporting to belong to news sources which

Address: 10 Brock Street
Regent Place
London, England
NW13FG

could mislead users. The group also meets once a week to update all members on the latest work being conducted by both platforms and publishers to combat misinformation. The initiative works as an information sharing exercise; there is no technical implementation by any party into each other's systems. We view the scheme as a valuable additional potential signal for misinformation on which we can, where appropriate, take action.

8. Given WhatsApp is an encrypted service, what plans does your company have to implement a reporting tool for content users receive within the app to you directly, which could allow your company to respond such as by withdrawing access to the app for repeat offenders or escalate reports and complaints to public authorities?

WhatsApp has a reporting tool which allows users to report message content directly to us, more information on which can be found here: https://faq.whatsapp.com/21197244/

We have also built advanced machine learning to help us detect accounts that are attempting to engage in bulk or automated messaging on WhatsApp, which is prohibited under WhatsApp's Terms of Service. We ban over 2 million accounts per month for attempting to abuse the service in this manner.

90% of all messages sent on WhatsApp are sent from just one WhatsApp user to one other WhatsApp user, and the average size of a group on WhatsApp is fewer than ten people. Further, WhatsApp does not have a search function or algorithms that would result in a particular piece of content being more or less likely to be viewed. As such, we take a different approach to tackling the spread of misinformation on WhatsApp.

Over the past two years, users have seen a steady drumbeat of changes to the WhatsApp service, all of which have been designed to constrain the virality of content on our service. We now label messages that have been forwarded with a single arrow in the top left-hand corner of the message. And when a message has been forwarded more than five times, we label it with a double arrow in the top left-hand corner of the message to signal to the user that the message they are looking at has been forwarded multiple times (we refer to these as "highly-forwarded messages", or HFMs). This helps our users to be more informed about the messages that they receive and share.

Last year, we reduced the maximum number of contacts that a user can forward on a WhatsApp message to from twenty to five. This resulted in a drop of 25% in

Address: 10 Brock Street
Regent Place
London, England
NW13FG

forwarding of messages globally, which translated to approximately one billion fewer messages being forwarded every day.

Last month, we went a step further and the limit on the number of contacts to whom a user can forward on a highly-forwarded message, from five chats to one. This has resulted in a 70% drop in this forwarding activity, although it is important to note that, even before we made this change, highly-forwarded messages only made up a very small percentage of the messages sent on WhatsApp.

9. Recently you set up the World Health Organisation with a WhatsApp account to be able to disseminate correct information easier. Is there scope to expand this to a limited number of authoritative partners, such as those in the Trusted News Initiative?

Yes—we are already working with further authoritative partners, including Public Health England in the UK, to roll out similar automated services that people can use to get the latest information. PHE's WhatsApp coronavirus information bot, which as I mentioned above has already sent more than one million messages, also provides a list of common hoaxes which is kept updated by the Government.

As you mention we partnered with the WHO on a new service, which is free to use, and has been designed to answer questions from the public about Coronavirus, and to give prompt, reliable and official information 24 hours a day, worldwide. This will also serve government decision-makers by providing the latest numbers and situation reports.

Just last week, we launched a new chatbot on WhatsApp in partnership with the International Fact-Checking Network. This bot connects WhatsApp users with independent fact-checkers in more than 70 countries, and with the IFCN's database of more than 7,500 debunked hoaxes related to the coronavirus. Using this bot, people from around the world are able to check whether a piece of content about COVID-19 has already been rated as false by professional fact-checkers.

This chat bot was made available as part of our $1m donation to the International Fact Checking Network to help address Coronavirus misinformation on all platforms. We have also provided an additional $1m dollar donation to the IFCN specifically to support more fact-checkers to begin to supplement their work using WhatsApp.

Address: 10 Brock Street
Regent Place
London, England
NW13FG

10. When will the new tool to show News Feed messages to people who have engaged with misinformation be rolled out in full to UK users? Will there be a phased release, i.e. will users in different territories have the tool rolled out at different times? Will a beta version be trialled beforehand?

These 'correct the record' messages started to roll out to all users, including users in the UK, on April 16th. They are displayed to people who have reacted to, shared, or commented on harmful misinformation about COVID-19 that we have since removed.

As with most messaging on Facebook, we began by testing a number of different options with small groups of users to measure which had the most impact, both in terms of click-through rate and in what they tell us about the messages' effectiveness. The messages have now been rolled out to 100% of users in all languages. The final messages can include the user's name and a specific description of whether they liked, commented, or shared the piece of content that was later removed.

These new messages relate to misinformation which has been removed from our platform because it can contribute to the risk of imminent violence or physical harm. They are separate to the notifications that we send to anyone who has previously shared a piece of content that is then rated as false by one of our third-party fact checkers. We launched this fact-checking programme in 2016.

11. Why does the tool only provide messages to users who engage with false content rather than those who see it, given your platform also measures signals such as impressions, reach, time spent on the post, click-through rate, video retention, etc, and thus will know who has seen misleading content without engaging with it?

We aim to strike a balance between two objectives here. On one hand, we want to show people who have engaged with misleading content either accurate information from a fact-checker or another authoritative source of information like the WHO mythbuster page, to give them more context in case they encounter that false information again either on our platforms or elsewhere. On the other hand, we do not want to draw attention to false narratives among people who may not have noticed them. We think the right balance is to show messages to those individuals who have activity engaged with the misleading content rather than those who simply saw it.

However, the 'correct the record' messages are only one part of our work to connect people on our services with accurate information. As described above, we signpost people looking for information about the coronavirus in the UK to either the NHS or

Address: 10 Brock Street
Regent Place
London, England
NW13FG

GOV.uk resources directly from our platforms, for example in search results for COVID-related terms or through our Coronavirus Information Centre at the top of people's News Feeds. We know that more than 350 million people have clicked through to these sources of information, with over 2 billion seeing these messages globally.

Off our platforms, we often run media literacy campaigns with partners to support people in thinking critically about the information they see. This includes Full Fact in the UK—during the 2019 UK General Election our joint media literacy campaign with Full Fact was seen by 11 million people and we are working with the UK Government to support a similar campaign focused on coronavirus.

12. Will this tool be backdated in any way, i.e. will a user who engaged with misinformation before the tool's roll out be notified of corrections through the tool after its roll out?

These 'correct the record' messages are displayed to people who have reacted to, shared, or commented on harmful misinformation about COVID-19 that we have since removed. We display those messages once the content has been removed to everyone who liked, reacted or commented on this harmful content within the previous seven days—even if that period included days prior to the tool being launched on April 16.

13. Will this tool only apply to misinformation and disinformation around the novel coronavirus and COVID-19, or will it be used to tackle misinformation and disinformation in general? What is the reason for this decision? If it is to be rolled out more broadly, how will misinformation and disinformation be identified?

The 'correct the record' messages currently only apply to misinformation around COVID-19, where there is widespread agreement from health organizations on the most harmful rumours and where we can direct people to resources like the WHO's list of debunked false claims. However, during several recent elections around the world, including the UK General Election in 2019, we showed similar messages to users who had seen misinformation that could have led to voter suppression, for example incorrect voting dates or methods.

Address: 10 Brock Street
Regent Place
London, England
NW13FG

14. How does Facebook define "harmful misinformation" and "imminent physical harm", and can you confirm that this extends to non-physical, indirect, vicarious or other forms of harm? Can you also specifically confirm that these examples are covered by your definition and clarify your logic when considering whether to take action:

a. the 'Stanford Hospital/St. George's Hospital medical advice' posts;
b. 5G conspiracy theory posts;
c. the 'St. Mary's bodybags' video; and
d. the 'crowded mosque' photo.

Within our Community Standards, Facebook uses the term "Misinformation that contributes to the risk of imminent violence or physical harm" to describe the type of misinformation that we remove from our platforms. We work with a wide variety of trusted partners to determine what types of misinformation this covers. In the context of coronavirus this has included working with health authorities—and the guidance from those outside experts has proven very helpful both for the teams who draft our Community Standards (the rules for what is and isn't allowed on Facebook and Instagram), and for our enforcement teams.

In the context of coronavirus, this type of misinformation includes:

- False information about the existence or severity of COVID-19, including, but not limited to: claims that COVID-19 does not exist, is not a pandemic, or that it is no more dangerous to people than the common cold or flu; and claims that COVID-19 government social distancing orders are a means of installing 5G wireless communication technology infrastructure or that the symptoms of COVID-19 are actually a result of 5G wireless technologies;

- False information about the means of preventing COVID-19 including, but not limited to: claims that something prevents someone from getting COVID-19 (e.g. existing vaccines, dietary practices, aromatherapy and essential oils); and misrepresentations of government guidance about the means for preventing the spread of COVID-19

- False information about how COVID-19 is transmitted, including, but not limited to: claims that any group is immune or cannot die from COVID-19 (e.g., children, people of certain races), or that a specific treatment or activity results in guaranteed immunity; and claims that 5G wireless communication technology causes the transmission of COVID-19.

- False information about cures, treatments, and tests for COVID-19, and false information about availability of essential services, including, but not limited to,

Address: 10 Brock Street
Regent Place
London, England
NW13FG

claims that essential services are now or soon to become unavailable, unless the appropriate governmental authority has publicly confirmed that information

These policies have been developed in consultation with outside experts. They do not use the terms “non-physical, indirect, or vicarious harms”, but we continue to talk with these experts about the ways different content on our platform might lead to harms, and how we should describe and take action on this content. Whenever we update our rules for what is not allowed on Facebook we make these changes public at on our website at www.facebook.com/communitystandards/recentupdates

In your question you list several different examples. The content and context of specific posts are essential to determining whether a piece of content breaches our Community Standards. In general, as demonstrated by the related Full Fact articles, most of these types of posts would be eligible to be reviewed by our third-party fact-checkers, but additionally our policies against harmful misinformation do cover false claims that 5G technology causes the symptoms of or contraction of COVID-19, which we remove from our platforms.

In other cases, where our fact-checkers rate content that contains information as false, we show it lower in people’s News Feeds so that fewer people see it, we cover it with a warning label where it does appear, and we give people who do see it (and anyone who shared it previously) additional context from our fact-checkers to help them get the full picture.

Thank you again for providing us with the opportunity to give evidence to your Committee, and for your letters. Please do not hesitate to get in touch again should you require further information as part of your inquiry.

Kind regards,

Richard Earley
UK Public Policy Manager

Address: 10 Brock Street
Regent Place
London, England
NW13FG

Google UK
Central St Giles
1 St Giles High St
London
WC2H 8AG

Julian Knight MP
Digital, Culture, Media and Sport Committee
House of Commons
London
SW1A 0AA

11th May 2020

Dear Mr Knight

Thank you very much for your letter dated 4 May following the committee to discuss COVID-19 misinformation.

Google is keen to participate fully in Parliamentary committees, answering questions and providing detailed clarity on our approach, policies, implementation and further areas of importance. I was pleased to be able to share with the committee new information on the work I have led in the UK including surfacing authoritative information from the NHS, partnering with the Government on promoting their public messages, supporting fact checkers and taking action against content that infringes YouTube's Community Guidelines. As with every committee, when questions arise which are outside of the main focus of the inquiry, or further information is requested, we are very pleased to follow up in writing.

1. Why has Google not implemented a report function for low quality or misleading search results, given the limitations of its automated systems and algorithms? How else can Google address these limitations?

Every search result page has a "send feedback" link at the bottom where users can provide exactly this sort of feedback, along with other observations or suggestions for how to improve. We have found that this kind of anecdotal reporting is not always the best way to address the important issues of low quality or misleading web pages in search results. Google indexes hundreds of billions of webpages and handles trillions of searches each year. Notably, every day, 15% of the queries we process are ones we've never seen before. Ensuring high quality results at that scale requires systemic approaches.

We continuously measure and assess the quality of our systems to ensure that we're achieving the right balance of information relevance and authoritativeness to maintain user trust in the results that they see. Google's algorithms identify signals about pages that correlate with trustworthiness and authoritativeness. We are constantly evolving these algorithms to improve results – not least because the web itself keeps changing. To that end, we continuously improve upon our systems using different mechanisms to capture feedback and continue to improve the quality of our results. These mechanisms include 1) Regular quality testing and evaluation and 2) Feedback mechanisms for reporting policy violations and illegal content, in line with applicable laws.

Regular quality testing and evaluation:
Google Search uses ranking algorithms to ensure we are meeting users' expectations of surfacing relevant and high quality sources. To help ensure Search algorithms meet high standards of relevance and quality, we have a rigorous process that involves both live tests and thousands of trained external Search Quality Raters from around the world. These Search Quality Raters must complete standard training and are tested on our rater guidelines before they can begin to provide ratings. These guidelines are publicly available and define our goals for Search algorithms -- including defining pages that potentially misinform users as "lowest quality."

Search Quality Raters assess whether a website provides users with the content they were looking for, and they evaluate the quality of results based on the expertise, authoritativeness, and trustworthiness of the content. These ratings do not directly impact ranking, but they do help us benchmark the quality of our results, which in turn allows us to build algorithms that globally recognize results that meet high-quality criteria. Search Quality Raters also perform evaluations of each improvement to Search we roll out: in side-by-side experiments, we show evaluators two different sets of Search results, one with the proposed change already implemented and one without. We ask them which results they prefer and why. This feedback is central to our launch decisions. In 2019 we ran over 464,065 experiments, with trained external Search Raters and live tests resulting in more than 3620 improvements to Search.

Feedback mechanisms for reporting policy violations and illegal content:
Google Search also contains some features, including knowledge panels, that are proactive in providing information. For these features, we have content policies to prevent content like hate speech, sexually explicit or misleading or inaccurate content from appearing. To that end, we provide more specific feedback channels alongside these types of features, in addition to the general feedback channel at the bottom of the Search results page that allow users to provide feedback on content that they believe is

in violation of our published content policies. We also provide an online form for requesting removal of content for legal reasons.

2. Do you accept the Global Disinformation Index's findings that Google has provided adverts for almost 90% of sites spreading coronavirus-related conspiracies? Can you point to where your policies have been updated publicly to ensure your advertising does not run on these types of websites?

Preserving the integrity of the ads and publishers on our platforms, as we're doing during the COVID-19 outbreak, is a continuation of the work we do every day to minimise content that violates our policies and stop malicious actors. We have thousands of people working across our teams to make sure we're protecting our users and enabling a safe ecosystem for advertisers and publishers. In 2019 alone, we terminated over 1.2 million accounts and removed ads from over 21 million web pages that are part of our publisher network for violating our policies. Terminating accounts—not just removing an individual ad or page—is an especially effective enforcement tool that we use if advertisers or publishers engage in egregious policy violations or have a history of violating policy.

Since the beginning of the COVID-19 outbreak, we've closely monitored advertiser behavior to protect users from ads looking to take advantage of the crisis. These often come from sophisticated actors attempting to evade our enforcement systems with advanced tactics. We have a dedicated COVID-19 task force that's been working around the clock. They have built new detection technology and have also improved our existing enforcement systems to stop bad actors. As per our policies, they are demonetising publisher content that includes claims about the propagation of COVID-19 that contradict the WHO or NHS guidance (such as theories involving 5G towers as a transmission vector). This is part of our existing Dangerous or Derogatory Content policy which prohibits monetization of content "promoting or advocating for harmful health or medical claims or practices".

These concerted efforts are working. We've blocked and removed tens of millions of coronavirus-related ads over the past few months for policy violations including price-gouging, capitalizing on global medical supply shortages, making misleading claims about cures and promoting illegitimate unemployment benefits.

On the specific research from Global Disinformation Index, the report does not explain what GDI defines as disinformation, nor does it provide the full list of domains sampled. So unfortunately, it is hard to peer review its findings. From what we gather, the revenue

estimates also do not accurately represent how publishers earn money on our advertising platforms.

3. What trends have you observed regarding the prevalence and targeting of gambling adverts during lockdown, and has this differed to previous trends? Are 'problem gamblers' ever targeted with gambling adverts and how do you protect them?

The safety of our users has always been our priority, and we do not target vulnerable users with any of our advertising tools. We require that any gambling advertisers using our services abide by local gambling laws and industry standards.

Beyond this, we restrict all gambling ads from showing to known minors, and only run display and YouTube ads to logged in over-18s. We do not allow gambling advertisers to use personalised advertising, nor do we allow the use of data collected by gambling advertisers to target their users on search with gambling ads.

On Search, gambling ads are not served against our gambling addiction related queries. This list consists of thousands of keywords and restricts gambling advertisers from targeting specific keywords such as "addiction". Our policy teams work to update this list on a regular basis. Google also provides a SafeSearch tool which can be utilised by those who have issues around gambling. With SafeSearch switched on, users will not see gambling ads.

4. How quickly does YouTube take action against content flagged as misleading? Has this response been expedited during the crisis?

We use a mix of machines and humans to enforce our policies at scale. Machines help us with scale and speed, whereas humans can bring judgement and can understand context. Our automated systems flag content that may be violative, helping us to more quickly and accurately enforce our Community Guidelines. Once potentially problematic content is flagged by our automated systems, human review verifies whether it indeed violates our policies. Our goal is to minimize views of violative content, and we've made significant strides in doing so. For example, between October and December 2019, we removed over 5.8 million videos from YouTube for violating our Community Guidelines. Of those removed videos, 90% were first flagged by machines, and 64.7% of those had never received a single view.

Because of the complicated and ever-changing nature of misinformation, we holistically approach misinformation through several policies in our Community Guidelines, including policies against spam, deceptive practices, scams, impersonation, hate, harassment, and content that encourages dangerous activities that could result in real-world harm. The policy in our Community Guidelines under which we've removed most violative coronavirus content is our "Harmful or dangerous content policy", which prohibits content promoting dangerous remedies or cures—e.g., videos that claim harmful substances or treatments can have health benefits. Context matters; each video is reviewed on its own merits and we are actively monitoring trends around COVID-19 content to ensure that our policies and enforcement evolve as the content does. We have removed thousands of videos promoting COVID misinformation from our platform, and the majority of these videos were viewed 100 times or fewer.

5. Regarding the London Real YouTube livestream of 5G conspiracy theories (the same interview was sanctioned by Ofcom for the radio broadcast by London Live), Google donated its video revenue to charity. However, the BBC reported that YouTube allowed London Real to keep its Super Chat revenue raised whilst the video was online. Can you clarify whether this was the case, and if so, why?

Our YouTube policies prohibit videos that promote medically unsubstantiated cures or treatments and we quickly remove flagged videos that violate these policies. Any content that disputes the existence or transmission of COVID-19, as described by the NHS and WHO, is in violation of YouTube policies. This includes conspiracy theories which claim that the symptoms are caused by 5G.

In line with this, the David Icke video which was referenced by the Committee has been removed due to violation of our policies. This action was taken, before any measures against the TV broadcast had been in place.

Users that repeatedly abuse the system and violate our policies can be penalised by being prevented from using livestream features, by having their channel demonetised, or by having their channel terminated. In line with this, David Icke's YouTube channel has been terminated.

In the case of the video referenced, following its removal, the revenue we received for Super Chats was donated to charity, however, the creator kept the revenue generated by this video while it was available.

As the situation evolves, so will our approach. That includes looking at further action as necessary to combat the threat of misinformation, and working with Government and industry to maintain the highest standards.

6. Facebook claimed that it only became aware of the London Real video after it was flagged to them by ITV. Do you think this is evidence that more cross-platform collaboration is needed in tackling misinformation? Did Google flag this through the Trusted News Initiative to other companies, and if not why not?

We believe collaboration with partners is the best way to ensure that everyone can get the information they need during this crisis, through whichever channel they use. The TNI is one of the forums where escalations happen and pieces of content are discussed among partners. The video in question was discussed in that forum after platforms had already taken action against it. We have also been working very closely with the Government, traditional media and, indeed, other tech companies, so that we can ensure that the overwhelming majority of what people find online is accurate and medically sound.

On March 16th, Google, YouTube, Facebook, LinkedIn, Microsoft, Reddit, and Twitter issued a joint statement saying “We are working closely together on COVID-19 response efforts. We’re helping millions of people stay connected while also jointly combating fraud and misinformation about the virus, elevating authoritative content on our platforms, and sharing critical updates in coordination with government healthcare agencies around the world. We invite other companies to join us as we work to keep our communities healthy and safe.”

7. How have you been working with partners in the Trusted News Initiative to ensure that authoritative information is surfaced appropriately and misinformation is demoted on your platforms?

As part of our efforts to collaborate with publishers and other platforms, we have continued to be an active and engaged partner within the Trust News Initiative - more recently focused on misinformation surrounding coronavirus, and previously on the UK election.

Partners within the initiative are able to share emerging trends, alert each other to misinformation about Coronavirus so that content can be reviewed promptly by platforms, whilst publishers ensure they don’t unwittingly republish misinformation.

Alerts also flag up content that undermines trust in partner news providers by identifying imposter content which claims to come from trusted brand identities or sources.

I want to stress that this partnership builds on our existing efforts to ensure authoritative information, including the work of fact checkers, is surfaced on our platforms. Google has highlighted fact checks in Search and News for over three years as a way to help people make more informed judgments about the content they encounter online.

These fact checks appear more than 11 million times a day in Search results globally and in Google News. That adds up to roughly 4 billion impressions a year. We have also created a library of over 40,000 fact checks which is publicly available for both users and journalists to consult through a dedicated search tool and for researchers to access through an open API.

8. Are you supporting or amplifying the output of traditional news organisations? If so, how, and if not, why not?

Throughout this crisis, we have sought to amplify the output of traditional news organisations and help direct users to their sites. Our business is built upon connecting users with authoritative, timely and relevant information. Never has this been more important than during the current crisis. We're committed to supporting the news ecosystem because we know high-quality journalism is vital during this difficult time.

To make it easier to find authoritative news, YouTube's Top News shelf prominently highlights videos from news sources in search results. And when a breaking news event happens, our Breaking News shelf highlights videos from news organizations about that event directly on the YouTube homepage. During the COVID-19 pandemic, we are displaying a Top News Shelf on YouTube search results pages for certain COVID-19 searches, and a Breaking News Shelf may also appear on the home page for users who are regular consumers of news content on YouTube. We are also displaying a dedicated COVID-19 news shelf to users in over 30 countries.

Throughout this period, we've seen a six and a half times increase in the amount of time people in the UK spend watching authoritative content and publishers such as The Guardian and The Telegraph have reached over one million YouTube subscribers for the first time.

In Google News, we've created a new COVID-19 section with up-to-date, relevant stories from the international to local level from a variety of authoritative sources. The section is now available to users across 25 top impacted markets, including the UK – and we continue to expand. The COVID-19 feature in Google News puts local news front and center with a dedicated section highlighting stories about the virus from local publishers in the reader's area. You can find more detailed information about these updates here.

We also recognise that the news industry is under increased pressure due to the economic impact of COVID-19, and at this critical time we are exploring additional ways we can continue to help support publishers and tackle fake news:

- The Google News Initiative has made a $6.5m investment into fact checking organisations. In the UK, we'll be supporting First Draft - one of the world's leading NGOs working on tackling disinformation - in their work to ensure journalists have the best training to be able to spot misinformation.
- To support local journalists throughout the crisis, the Google News Initiative has also launched a global Journalism Emergency Relief Fund, which will support small- and medium-sized news organisations producing original news for local communities.
- We've provided fee relief to support our news partners, as well as a $1 million Google.org donation to the International Center for Journalists, providing immediate resources to support reporters globally.

9. On 23 April, the US President asked the Coronavirus Response Coordinator to test exposure to ultraviolet light and injections of disinfectant on patients. The White House streams its daily briefings to YouTube. Would YouTube therefore take action against the White House's video if the claims were streamed unchallenged, given YouTube's statement that it will ban conspiracy theories from the site?

At YouTube, we enforce our Community Guidelines across the board, regardless of who is the content creator. The application of these policies is always a function of the content at stake and of the context in which it was posted. For instance, educational or journalistic organisations' reasons for posting potentially problematic content are very different from those users who'd be less well intentioned.

It's difficult to make general statements about hypothetical videos, as the exact nature and context of the content would be of paramount importance for a determination.

However, we have not taken action against the White House video that you mention. It is important to note that the video included context provided by others who spoke during the conference, for example, William Bryan, Undersecretary for Science and Technology at the Department of Homeland Security.

Thank you again for the opportunity to share these details about our fight against disinformation. Please do not hesitate to get in touch if we can help with anything else.

Yours sincerely,

Alina Dimofte
Government Affairs and Public Policy Manager

Twitter UK
20 Air Street
London
W1B 5AN

twitter.com

11 May 2020

Dear Chair,

Thank you for your letter.

We regret the Committee's concerns as outlined, and welcome the opportunity to address them. The format was challenging and, for future hearings, we would be pleased to appear individually - which may allow more time to cover these questions. We did raise this issue with the Committee Specialist in advance, and would be pleased to explore an alternative approach in future.

We have provided a detailed response attached, which includes both answers to your questions, as well as the positive trends and steps taken that we did not cover during the hearing.

If you have any further questions, we would be happy to meet directly.

Please see attached:

- **Our Covid-19 misinformation strategy in the UK**
- **Answers to your questions**
- **Retrospective Review of the 2019 General Election**
- **Transcript from the Democracy and Digital Technologies Select Committee, March 2020**
- **Timeline of updates 2014-2020**

Yours sincerely,

Katy

Covid-19 misinformation strategy in the UK

Below is a summary of work undertaken in the UK. It is by no means comprehensive, and so please do let us know if the Committee has any further questions.

We share information about our global strategy on an ongoing basis via our Covid-19 blog, and through our partnerships in the UK with government and civil society. As the Culture Secretary stated in his testimony to the Committee in April, the 'single best thing' we can do is to drive reliance on reliable narratives. Our entire company is focused on ensuring Twitter surfaces authoritative information on Covid-19.

The power of a uniquely open service during a public health emergency is clear. The speed and borderless nature of Twitter presents an extraordinary opportunity to get the word out and ensure people have access to the latest information from expert sources around the world. To support that mission, our global Trust & Safety team is continuing its zero-tolerance approach to platform manipulation and any other attempts to abuse our service at this critical juncture.

Policy

- On March 18, we shared the 11-point criteria our team would be using to assess whether Tweets were considered harmful Covid-19 misinformation and against our rules.
- We broadened our guidance in April on unverified claims that incite people to engage in harmful activity, could lead to the destruction or damage of critical 5G infrastructure, or could lead to widespread panic, social unrest, or large-scale disorder.
- Since introducing our updated policies on March 18, we shared on April 22 that we had removed over 2,230 Tweets containing misleading and potentially harmful content.
- We continue to advise on our website that users should follow @WHO and their local health ministry; and they should seek out the authoritative health information and ignore the noise. We ask them to report it to us immediately if they see something suspicious or abusive. Most importantly, we advise users to think before you Tweet.

Partnerships

- Now, over 40 UK government partners and health authorities are onboarded to our Partner Support Portal, which enables trusted partners to expedite a wide range of issues to a dedicated team within Twitter.
- In April, we provided the government with pro bono advertising space and support to promote key #StayHomeSaveLives public health messaging. Not only has this had significant reach, but we also saw people actively participating - committing to #StayHomeSaveLives and the key behaviours required to keep us all safe.
- We also continue to provide pro bono advertising space and support to a number of charities, including around fact-checking and digital literacy. More broadly, for educators

and parents, we have a specific media literacy guide, which was created in partnership with UNESCO. We provided this to Ofcom to share with their wider Making Sense of Media Network. In April, we ran a campaign with UNESCO and the EU in promoting media and information literacy. Throughout the week they shared best in class ways to critically analyse what we engage with online and encourage people to #ThinkBeforeSharing.

Product

- Launched 6 days before the official designation of the virus in January, we partnered with the Department of Health and Social Care to launch a dedicated prompt feature at the top of Search result. This ensures that when you come to the service for information about Covid-19, you are directed to the NHS website where you are met with credible, authoritative content. In May, we developed this further in partnership with the Department for Digital, Culture, Media and Sport - such that when people search for 5G information on Twitter, they are served with a prompt encouraging them to 'Know the facts'; stating that the UK Government has seen no link between 5G and Covid-19; and directing to a new Gov.uk resource.

- Twitter Moments are a curated series of Tweets to tell a story. We have ensured a live, up-to-date Moment, with the latest Tweets from the key UK government and health accounts about Covid-19, is available at the top of the home timeline for all UK users. We have seen a 45% increase in usage of these pages globally.

- We shared on April 22 that our automated systems have challenged more than 3.4 million accounts targeting manipulative discussions around Covid-19.

Research

- Twitter is the only major service that makes public conversations available for study. Tens of thousands of researchers have used our public API (technical interface to allow access to our public data) over the past decade.
- To provide transparency on the Covid-19 conversation on Twitter and further support the research community, on April 30 we launched a new program that offers free access to a dedicated Covid-19 API endpoint. In practice, this will allow approved developers and researchers to access public conversations about Covid-19 across languages, resulting in a data set that will include tens of millions of Tweets daily. The data can be used to research a range of topics related to the coronavirus pandemic, including areas like the spread of the disease, the spread of misinformation, crisis management within communities and more.
- Twitter's public archive of state-backed information operations is the largest of its kind in the industry. We continue to publicly disclose and make datasets available for research. First launched in October 2018, and continuously updated, the archive has been accessed by thousands of researchers from around the world, who in turn have conducted independent, third-party investigations of their own. In total, there are now 28 datasets, comprising 8 Terabytes of data.

Answers to questions

1. What proportion of identified accounts spreading misinformation and disinformation are bots or automated accounts or use automation in some way (irrespective of the number of overall/legitimate accounts using automated functions like scheduling)?

Automated accounts themselves are permitted as long as they do not break the Twitter rules. Bots, or fully automated accounts, can be a positive and vital tool, from customer support to public safety. We strictly prohibit, however, the use of bots and other networks of manipulation to undermine the core functionality of our service. As I said in the hearing, on Covid-19 specifically we shared on 22 of April that we had challenged 3.4 million spammy or manipulative accounts. More broadly, as we shared in our last Transparency Report, we challenged 97 million accounts spammy or manipulative in the first 6 months of 2019.

We do not identify and share specifically the number of automated accounts in the way we do with violations of our rules in our Transparency Reports. Automation is used frequently in day-to-day social media management - it would be challenging to compile a specific number of automated accounts at any one time.

We do, however, provide Tweet source labels to help people better understand how a Tweet was posted, since Tweets and campaigns can be directly automated by an application. At the bottom of a Tweet, you'll see the label for the source of the account's Tweet. For example, "Twitter for iPhone," "Twitter for Android," or "Twitter for Web." In some cases you may see a third-party client name, which indicates the Tweet came from a non-Twitter application. Authors sometimes use third-party client applications to manage their Tweets, manage marketing campaigns, measure advertising performance, provide customer support, and to target certain groups of people to advertise to.

2. If verification is simply validation of identity, why does Twitter not roll out verification to all users who opt in (leaving the option of anonymity or pseudonymity for those who choose not to) to tackle disinformation, abuse and other harms?

When Twitter was founded, we offered a range of ways for people to manage their privacy and control their experience on Twitter - from creating pseudonymous accounts, to letting people control who sees their Tweets, and a wide array of granular privacy controls.

This has to be balanced by our rules, and the information we ask for from users when they signup to Twitter. Upon signup, we ask users for their full name and email address or phone number. Relevant rules include (but are not limited to):

- You can't mislead others on Twitter by operating fake accounts;

- You can't artificially amplify or disrupt conversations through the use of multiple accounts;
- You may not use Twitter's services in a manner intended to artificially amplify or suppress information or engage in behavior that manipulates or disrupts people's experience on Twitter.

We enforce these rules through a range of reactive and proactive measures, as listed above.

The blue verified badge on Twitter lets people know that an account of public interest is authentic.

Verification was meant to authenticate identity and voice, but it has also been interpreted as an endorsement or an indicator of importance. With that in mind, we closed all public submissions for verification in November 2017. We are continuing to review our verification programme, and will keep the Committee updated.

3. Have you taken specific action against the accounts of politicians, celebrities, influencers, etc, such as by removing verified status or deleting accounts? What action is typically taken and are these accounts notified about why action is taken?

While we avoid providing public commentary on specific individual accounts, we have removed Tweets that break our rules. This includes from accounts that are verified. It has also included using our public interest interstitial. Our full range of enforcement actions is available here.

Critically, given Twitter's public nature, the Committee can look at an account or Tweet on the service at any time to see if we have taken action for violating our rules. If a Tweet has been removed or an account has been suspended, it will say so.

4. What defensive measures are used, such as how you identify specific tweets, hashtags and retweets, bots or automated accounts and hostile foreign actors that spread mis/disinformation, and how you disrupt or remove this activity? (Please provide confidential information in a supplementary response in the interest of transparency.)

See attached supplementary response.

5. Facebook specifically allows users to flag posts specifically as 'false news'; Twitter allows users to report fake accounts but not false news specifically. Does your company have any plans to roll out functionality for users to specifically report posts as misleading, and if not why not?

We are exploring a range of approaches to tackling misinformation.

As you state, users can already report spam, suspicious and fake accounts and Tweets directly, as well as Tweets that contain links to harmful sites (and indeed a wide range of violations of our rules).

Our aim with harmful Covid-19 misinformation is to rapidly identify and remove Tweets that pose the greatest risk of causing harm. User reports are not the only way to identify content - we have technical signals, machine learning models, expert partnerships and government reports. Often user reports can add a lot of noise to the system, slowing down response and enabling people to report Tweets with which they simply disagree - not because they break the rules.

I am attaching our Retrospective Review of the 2019 General Election as an example. We launched a user reporting tool for voter misinformation over the campaign; we found, however, that the majority of Tweets we removed for breaking our rules on voter misinformation were detected proactively through our own systems, and that user reports were a far less effective indicator of urgency and priority.

6. Twitter's spokesperson, Katie Rosborough, has previously stated that the company "will not take enforcement action on every Tweet that contains incomplete or disputed information about Covid-19".1 What is the reason for this approach, given that it implies that Twitter will not take action against disputed information even where you are aware of it?

That is correct. The information environment and what is accurate is evolving rapidly, with different and disputed advice from governments all over the world, and even varying approaches from local or regional, as well as national, governments. What's more, there is a public interest in enabling conversations with certain types of disputed disinformation - critically, researchers are major users of Twitter, and conversation may involve someone discussing a research study, or professionals exploring potential avenues for treatment. People also need to be able question the efficacy or proportionality of measures enacted by governments and other authorities. The policy debate between the Government, the opposition, the media and experts highlights that there is no single correct answer, in the vast majority of situations - so a statement claiming that a company will remove every disputed piece of information would require that company to adjudicate on every statement to decide if it is accurate.

That is very different to Tweets with harmful misinformation regarding Covid-19, that fall into one of the 11 categories detailed on our website and below.

An update on our content moderation work

We have broadened our definition of harm to address content that goes directly against guidance from authoritative sources of global and local public health information. Rather than reports, we are enforcing this in close coordination with trusted partners, including public health

authorities and governments, and continue to use and consult with information from those sources when reviewing content.

Under this guidance, we will require people to remove Tweets that include:

- Denial of global or local health authority recommendations to decrease someone's likelihood of exposure to Covid-19 with the intent to influence people into acting against recommended guidance, such as: "social distancing is not effective", or actively encouraging people to not socially distance themselves in areas known to be impacted by Covid-19 where such measures have been recommended by the relevant authorities.
- Description of alleged cures for. Covid-19, which are not immediately harmful but are known to be ineffective, are not applicable to the Covid-19 context, or are being shared with the intent to mislead others, even if made in jest, such as "coronavirus is not heat-resistant — walking outside is enough to disinfect you" or "use aromatherapy and essential oils to cure Covid-19."
- Description of harmful treatments or protection measures which are known to be ineffective, do not apply to Covid-19, or are being shared out of context to mislead people, even if made in jest, such as "drinking bleach and ingesting colloidal silver will cure Covid-19."
- Denial of established scientific facts about transmission during the incubation period or transmission guidance from global and local health authorities, such as "Covid-19 does not infect children because we haven't seen any cases of children being sick."
- Specific claims around Covid-19 information that intends to manipulate people into certain behavior for the gain of a third party with a call to action within the claim, such as "coronavirus is a fraud and not real — go out and patronize your local bar!!" or "the news about washing your hands is propaganda for soap companies, stop washing your hands."
- Specific and unverified claims that incite people to action and cause widespread panic, social unrest or large-scale disorder, such as "The National Guard just announced that no more shipments of food will be arriving for two months — run to the grocery store ASAP and buy everything!"
- Specific and unverified claims made by people impersonating a government or health official or organization such as a parody account of an Italian health official stating that the country's quarantine is over.
- Propagating false or misleading information around Covid-19 diagnostic criteria or procedures such as "if you can hold your breath for 10 seconds, you do not have coronavirus."
- False or misleading claims on how to differentiate between Covid-19 and a different disease, and if that information attempts to definitively diagnose someone, such as "if you have a wet cough, it's not coronavirus — but a dry cough is" or "you'll feel like you're drowning in snot if you have coronavirus — it's not a normal runny nose."

- Claims that specific groups or nationalities are never susceptible to Covid-19, such as "people with dark skin are immune to Covid-19 due to melanin production" or "reading the Quran will make an individual immune to Covid-19."
- Claims that specific groups or nationalities are more susceptible to Covid-19, such as "avoid businesses owned by Chinese people as they are more likely to have Covid-19."

7. Assuming the proposed duty of care set out in the Online Harms White Paper and initial consultation response has public and civil society buy-in, would Twitter support it in principle, given that the company has said that it will not take enforcement action on every tweet as per the above?

As I stated in another House of Commons Select Committee last year, "From our perspective, there are plenty of really positive aspects of the White Paper - not least the creation of a new regulator, which could be a big step forward."[1] As our Chief Executive has stated, we are in favour of smart regulation. Most importantly of all, however, we are not waiting for regulation. Attached is a timeline of the changes we have made, including in the period since Online Harms was first proposed back in 2017.

Critically, the Government has so far only published an initial response to the Online Harms White Paper. They noted that they were minded to appoint Ofcom as the new regulator, which was something we supported in our White Paper submission. Once the Government publishes their full response and/or a draft Bill, we will certainly review and can provide our feedback.

This area does highlight the challenge in reconciling a Duty of Care approach with tackling misinformation. The very nature of debate in a free society is that people debate issues, dispute evidence and disagree over what is true. This happens every day in Parliament. That debate and disagreement is ever present on Twitter, happening in public, all around the world. If a regulator expected every piece of false information to be removed from a service to fulfil a Duty of Care, that would place companies - and ultimately the regulator - in a position of evaluating every piece of content for truthfulness.

We have had wider discussions about UK regulation in other Committees (such as the House of Lords Select Committee on Democracy and Digital Technologies in March, transcript attached), and would be very happy to discuss this with the Committee further.

8. Ms. Rosborough's statement claims that "[w]e're prioritizing the removal of content when it has a call to action that could potentially cause harm". Can you be as specific and exhaustive as possible about what Twitter considers constituting a "call to action" and "potentially [causing] harm"? How is illegal or harmful content that does not specifically include a call to action prioritised?

[1] Home Affairs Select Committee, 24 April 2019

Yes. In our answer to question 6, we provide the full breakdown of our policy. Examples of a call to action, for instance, would include “coronavirus is a fraud and not real — go out and patronize your local bar!!” or “the news about washing your hands is propaganda for soap companies, stop washing your hands.” Our harmful misinformation policy for Covid-19, however, is not limited to calls to action. It also includes a wide range of other harmful content that may not include a specific call to action. For instance, description of alleged cures for Covid-19, which are not immediately harmful but are known to be ineffective; and denial of established scientific facts about transmission during the incubation period or transmission guidance from global and local health authorities, such as “Covid-19 does not infect children because we haven’t seen any cases of children being sick.”

We evaluate whether content is against our rules. For potentially illegal content, police can contact us any time through Twitter’s Legal Online Submissions Site - which is a dedicated reporting channel for law enforcement with 24/7 coverage - to enable police investigation and ultimately the courts to decide if the law has been broken.

9. Does Twitter today have the same number of full-time equivalent staff working directly in content moderation as before social distancing measures were in place (i.e. 1 February 2020), both for English-language users and globally?

Our priority is the health and safety of our staff. With that in mind, on 11 March we informed all employees globally they must work from home. For contractors and hourly workers who are not able to perform their responsibilities from home, Twitter will continue to pay their labor costs to cover standard working hours while these restrictions are in effect.

We are not immune from impact. We are providing regular updates on our website on what you can expect if you make a report to Twitter. To further protect the conversation, we are:

- Instituting a global content severity triage system so we are prioritizing the potential rule violations that present the biggest risk of harm and reducing the burden on people to report them.
- Executing daily quality assurance checks on our content enforcement processes to ensure we’re agile in responding to this rapidly evolving, global disease outbreak.
- Engaging with our partners around the world to ensure escalation paths remain open and urgent cases can be brought to our attention.
- Continuing to review the Twitter Rules in the context of Covid-19 and considering ways in which they may need to evolve to account for new behaviors.

We are also increasingly using technology:

- To help us review reports more efficiently by surfacing content that's most likely to cause harm and should be reviewed first.

- To help us proactively identify rule-breaking content before it's reported. Our systems learn from past decisions by our review teams, so over time, the technology is able to help us rank content or challenge accounts automatically.
- For content that requires additional context, such as misleading information around Covid-19, our teams will continue to review those reports manually.

10. How are you balancing moderation of different kinds of harmful material? Are there some categories that are being expedited?

In the UK, we have over 40 government partners able to make a full range of reports to us via our Partner Support Portal, which are expedited to a dedicated team within Twitter.

We continue to speak regularly with the Home Office and NCA. We have not recommended any adjustments to reporting processes for law enforcement at this time. We continue to use proactive technologies to identify violations of our rules, particularly in surfacing the most egregious content - indeed, when it comes to terrorism and child sexual exploitation, a vast majority of the accounts that break these rules have been identified proactively for some time.

Retrospective Review

Twitter and the 2019 UK General Election

March 2020

Table of Contents

Foreword

Twitter's purpose is to serve the public conversation, particularly around election cycles. We want to encourage more healthy, open political debate, and to strengthen civic participation by providing a platform for people to engage in conversations that impact their everyday lives.

The fight against malicious activity and abuse goes beyond any single election or event – it's ongoing. Enhancing our safety policies, developing better tools and resources for finding and stopping abuse, and taking extensive action against activity that violates the Twitter Rules is just some of what we do every day.

We work alongside political parties, researchers, experts, and election commissions and regulators around the world, all while investing in our proactive detection and enforcement efforts on the platform. We also stay in touch with national parties and state election officials on an ongoing basis to be sure they know how to report suspicious activity, abuse, and rule violations to us.

Ahead of the UK General Election 2019, we established a cross-functional team to proactively protect the integrity of the election-related conversation on our platform and maximise Twitter as the go-to place for people to see what's happening, participate in the conversation, and track the campaign trail. As part of our election integrity efforts, we also worked to support partner escalations and to identify potential threats from malicious actors. At all times we strove to be transparent - sharing the steps we took publicly during the General Election on our blog (here and here), and openly engaging with partners in government, industry and civil society throughout. This document is a summary of the work we did, and the key results and learnings we achieved as a result of this collaborative process.

Katy Minshall
Head of UK Government, Public Policy and Philanthropy

Key figures

- **More than 15 million Tweets were posted about the election** - highlighting how political discourse on the service continues to thrive.
- Compared with the UK election in 2017, there was a **66% increase in the volume of the Twitter conversation from the beginning of the campaign to the end of election day.**
- Approximately **170,000 people in the UK clicked through to Twitter's custom prompt,** created in collaboration with Democracy Club, to find out where their local polling station was and who was on their ballot.
- The activating emoji hashtags of; #GE2019, #GE19, #GeneralElection19 and #General Election 2019 were used over **6 million times** — showing the power of emojis and hashtags in bringing the conversation together throughout this election.
- **Overall, the highest volume of Tweets we saw was related to the exit poll results. At 10:05pm, there was an influx of almost 10,000 Tweets posted over 60 seconds.**
- **The 9 live streams from news partners over election night totalled 3.1 million video views.**
- We took action on **2,073 Tweets** from our proactive spam monitoring tools over the duration of the UK General Election (from November 11th to December 19th).
- From reports made via our public-facing voter misinformation reporting tool, we took action on **167 cases of election-related misinformation. Our team also took proactive action on a total of 406 Tweets** for violating our voter misinformation policy on election day.
- As part of our effort to surface credible election information during the campaign, we **curated 121 Moments and Events** related to the election, with 8 of these focused specifically on debunking misleading information. Users visited and engaged with these Moments and Events millions of times over the course of the campaign.
- **We worked with the Home Office and peer companies to provide information to all 43 police forces,** and a dedicated 'one stop shop' for candidates on safety advice across all social media platforms.
- We received **497 reports from partners on the Partner Support Portal during the election, with an action rate of 93.3%.** We now have almost 40 UK partners on our Partner Support Portal, able to escalate a wide range of issues to a dedicated team within Twitter.

Civic participation, and following the election on Twitter

During any election, Twitter is the place for people to see what's happening, participate in the conversation, and track the campaign trail. Throughout the six weeks of campaigning, there was a consistently high level of political debate on UK Twitter, including from candidates, parties, voters, journalists, civil society groups, and interested election-watchers.

- To unite people around the UK election conversation on Twitter, we launched a custom election emoji that ran for the duration of the campaign.
- To encourage civic participation, we launched a prompt created in collaboration with Democracy Club, to encourage users to find out where their local polling station was and who was on their ballot. This ran twice - from 3 days before the election; and on election day itself (see right).
- We worked with 18 news publishers and broadcasters to ensure over 300 journalists were up-to-date on best practice.
- We worked with every major broadcaster in the UK to ensure there was video content from every debate in the lead up to the election (several were live-steamed; others produced as-live clips) - with all partners sharing a link to Twitter's event pages ahead of time to drive users to set reminders. ITV's flagship political show, Peston, live streamed an interview with Boris Johnson and Nicola Sturgeon on Twitter 3 hours before it was broadcast on TV.
- On election day, we activated a custom emoji to support conversation around #DogsAtPollingStations.

Highlights:

- **More than 15 million Tweets were posted about the election** - highlighting how political discourse on the service continues to thrive. Compared with the UK election in 2017, there was a **66% increase in the volume of the Twitter conversation from the beginning of the campaign to the end of election day.**
- Approximately **170,000 people in the UK clicked through to Twitter's custom prompt,** created in collaboration with Democracy Club, to find out where their local polling station was and who was on their ballot.

- The activating emoji hashtags of; #GE2019, #GE19, #GeneralElection19 and #General Election 2019 were used over **6 million times** — showing the power of these hashtags in bringing the conversation together throughout this election.
- **Overall, the highest volume of Tweets we saw was related to the exit poll results. At 10:05pm, there was an influx of almost 10,000 Tweets posted over 60 seconds.**
- **The 9 live streams from news partners over election night totalled 3.1 million video views.**
- Conversational volume relating to both Conservative and Labour leaders was neck-in-neck for the last two weeks of the campaign, up until 24 hours before election day — with only a 0.32% difference in Tweet volume between the two at that stage.
- We saw over 122,000 #DogsAtPollingStations Tweets throughout #ElectionDay.

Proactive Improvements

Learning from previous elections and tactics used by malicious actors, we engaged in intensive efforts to identify and combat attempts to undermine and abuse our service. Ahead of the General Election, we took a number of proactive steps to protect and support the election conversation:

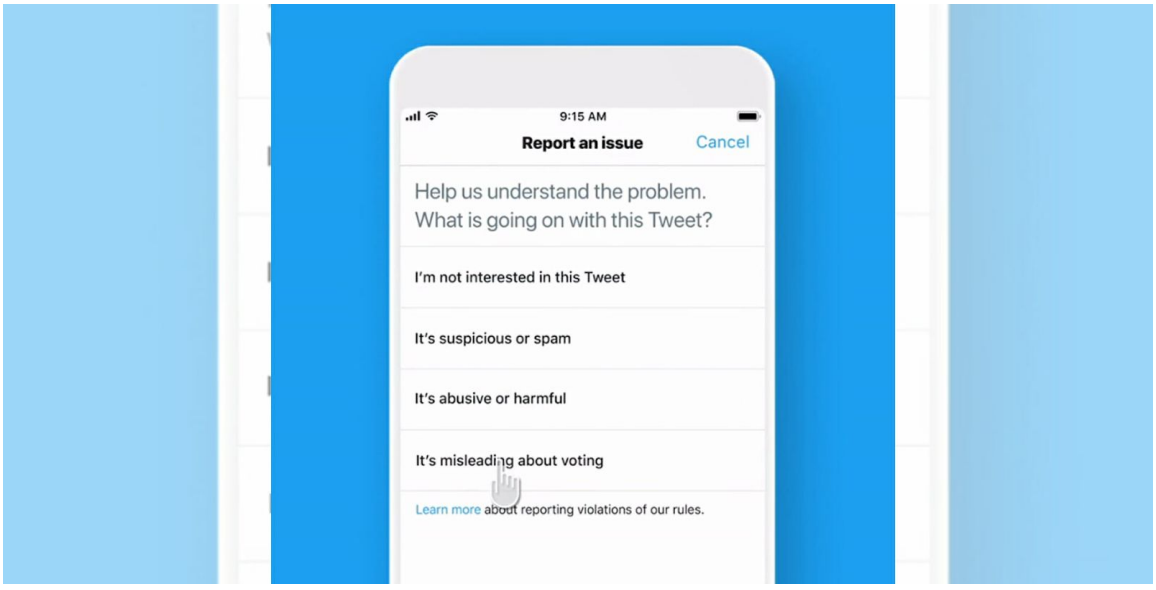
- Announcing a prohibition on political advertising, and enforcing it from 22 November 2019.
- Introducing a new tool for the UK election which allowed citizens to report deliberately misleading election-related content to us.
- Launching a public consultation process around our new synthetic and manipulated media policy. We subsequently announced the results of this consultation and our policy on 4 February 2020, and communicated this directly with UK government stakeholders.

We continued to engage on a regular basis with DCMS and other government stakeholders during the course of the election.

Highlights:

.

- We took action on **2,073 Tweets** from our proactive spam monitoring tools over the duration of the UK General Election (from November 11th to December 19th). Our rules on platform manipulation and spam policy are available here.
- From reports made via our public-facing voter misinformation reporting tool (see below), we took action on **167 cases of election-related misinformation. Our team also took proactive action on a total of 406** Tweets for violating our voter misinformation policy on election day.



- Having participated in the BBC's Trusted News Initiative for the UK General Election, we received no reports from them throughout the campaign.
- As part of our effort to surface credible election information during the campaign, we curated **121 Moments and Events** related to the election, with 8 of these focused specifically on

debunking misleading information. Users visited and engaged with these Moments and Events millions of times over the course of the campaign.

Safety

We want people to feel safe on Twitter, and we are continually striving to be more proactive and open in the work we do to serve the public conversation on Twitter. More than 50% of Tweets we take action on for abuse are now proactively surfaced using technology, rather than relying on reports to Twitter; and we have seen a 105% increase in accounts actioned by Twitter (locked or suspended for violating the Twitter Rules). During the election campaign, we took a number of steps to enable candidates, political parties and campaigners to feel safe on Twitter.

- We created bespoke safety and best practice resources for candidates, which were disseminated nationally.
- **We worked with the Home Office and peer companies to provide information to all 43 police forces,** and a dedicated '[one stop shop]' for candidates on safety advice across all social media platforms.
- We promoted among all MPs and major political parties our dedicated channel for candidates and campaigners to get in touch, and worked with government departments to further promote our safety advice and resources.
- We provided 3 safety and security webinar trainings for candidates, covering best practice, safety tools such as mute, block and report, and security tips and techniques (below).

Highlights:

- We received **497 reports from partners on the Partner Support Portal during the election, with an action rate of 93.3%.** We now have almost 40 UK partners on our Partner Support Portal, able to escalate a wide range of issues to a dedicated team within Twitter.
- We continued to speak on a weekly basis with the Parliamentary Security Department.
- We received less than 40 queries via our dedicated govuk@twitter.com channel during the campaign. The vast majority of these were regarding attending our webinars for candidates; the remainder were mostly questions about verification and changes to usernames.

Conclusion

As with every election around the globe, election integrity and online safety are key priorities for Twitter, and #GE2019 was no different. To this end, our strategy was to ensure the conversation was healthy, open, and safe for everyone and we implemented a clear plan to safeguard the UK election.

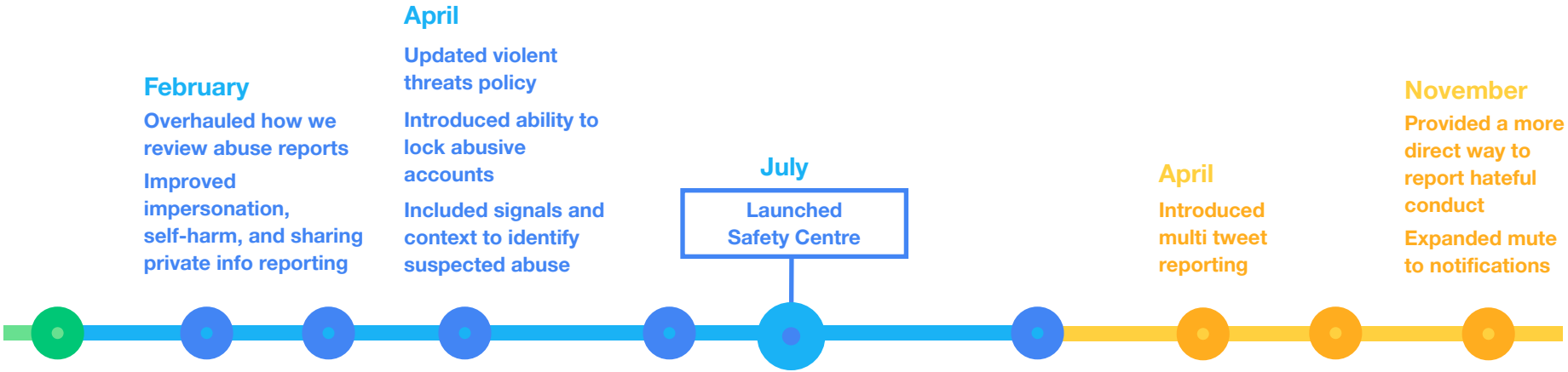
Our global cross-functional teams worked proactively to protect the integrity of regional trends, support partner escalations, and identify potential threats from malicious actors. Our teams continue to review and assess potential malicious or automated activity by bad actors, either foreign or domestic, and when and where we identify evidence of state-backed information operations on Twitter, we routinely publish archives of Tweets and media associated with those operations. We do so to encourage researchers to conduct their own investigations, and to share their insights and independent analysis with the world.

We'll continue supporting and protecting conversations on the service, particularly during election cycles, and we'll strive to be open and transparent as we enhance and refine our approach.

Safety Updates

2014 - 2020

Timeline



December
Improved harassment reporting

Enhanced block function

February
Overhauled how we review abuse reports

Improved impersonation, self-harm, and sharing private info reporting

March
Made it easier to report threats to Law Enforcement

April
Updated violent threats policy

Introduced ability to lock abusive accounts

Included signals and context to identify suspected abuse

June
Launched block lists

July
Launched Safety Centre

December
Updated Twitter Rules on abusive behavior and hateful conduct

April
Introduced multi tweet reporting

June
Enabled users to block directly from a tweet

November
Provided a more direct way to report hateful conduct

Expanded mute to notifications

2014 2015 2016


jack
@jack
We're taking a completely new approach to abuse on Twitter. Including having a more open & real-time dialogue about it every step of the way
527
1.4K

Timeline

March

- Leveraged technology to reduce abusive content
- Expanded mute and notification filters
- Began sending in-app report notifications

October

- Updated Non-Consensual Nudity Policy
- Updated Private Information Policy
- Improved communication around appeal process

December

- Implemented account relationship signals to improve witness reporting
- Specifying policy in violation to witnesses who submit reports
- Added new Policy that does not allow hateful display names

February

- Enabled users to report tweets even if blocked
- Stopped the creation of new abusive accounts
- Introduced safe search functionality
- Began collapsing potentially abusive or low-quality tweets

June

- Held first Trust & Safety Council Summit

August

- Updated Safety Center

November

- Sharing offending tweet and explanation for locked accounts
- Updated Twitter Rules on self-harm, spam & related behaviors, and graphic content & adult violence
- Updated Verification Policy on loss of status

2017

- Updated Media Policy to not permit hateful imagery and hate symbols in profile elements
- Began suspending violent groups accounts and removing content that glorifies or condones acts of violence
- Expanded unwanted sexual advances enforcement by integrating relationship interaction signals

Timeline

February

Users can report content that encourages self-harm or suicide

March

Twitter announces Request for Proposal to assist with Health initiative

Users suspended for abusive behaviour emailed with violating content & broken rule

Users can opt-in to get Tweets included in report receipts, in-app and by email

April

Report flow is updated to better define 'protected category'

May

Behaviour-based signals introduced to influence how Tweets appear in Search & Conversation

June

Acquisition of Smyte, a company that specializes in safety, spam, and security issues

Users can avail of security keys for login verification

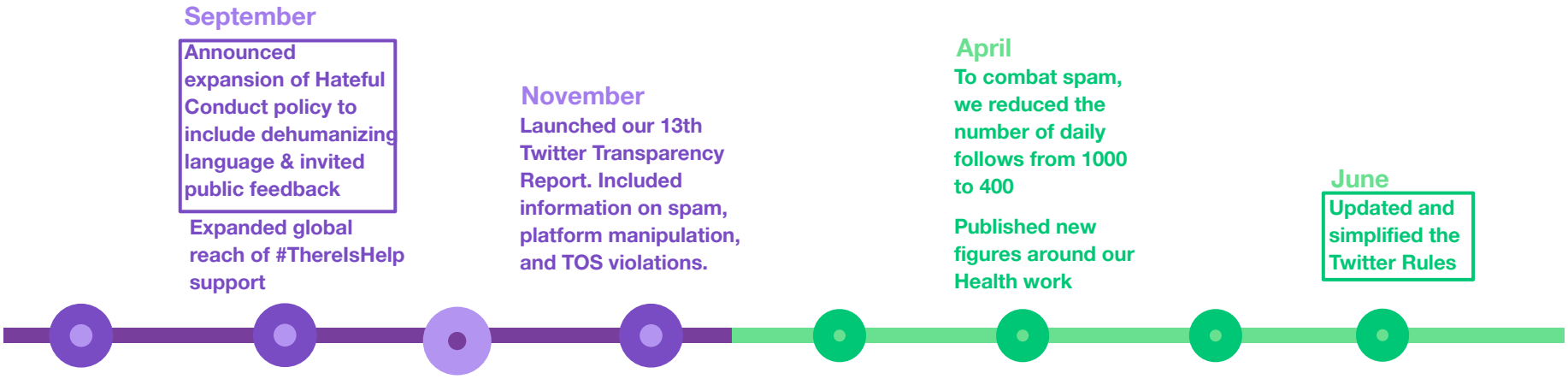
July

More stringent enforcement of the policies related to messages sent during live broadcasts

Twitter to work with Universities of Oxford and Leiden to measure Health work

2018

Timeline



2018

2019

Timeline

June
Introduced a notice to provide more context and clarity around enforcement of Tweets that are in the public interest.

July
Updated our rules against hateful conduct to include language that dehumanizes others on the basis of religion.

September
Announced a new policy to address financial scams on Twitter.

November
Launched our 15th Twitter Transparency Report, which shows a 105% increase in Twitter's proactive enforcement on accounts.

Provided users globally with the ability to moderate replies.

December
Announced we are strengthening our Trust & Safety Council of issue experts to include more diverse voices and foster deeper conversations.

New disclosures on state-backed information operations.

February
Introduced a new policy on synthetic and manipulated media after consulting with civil society and academic experts.

2019

2020

Select Committee on Democracy and Digital Technologies

Uncorrected oral evidence: Democracy and Digital Technologies

Tuesday 17 March 2020

2.35 pm

Watch the meeting

Members present: Lord Puttnam (The Chair); Lord German; Lord Harris of Haringey; Lord Holmes of Richmond; Baroness Kidron; Lord Lipsey; Lord Lucas; Baroness McGregor-Smith; Baroness Morris of Yardley.

Evidence Session No. 25 Heard in Public Questions 311 - 325

Witnesses

I: Katy Minshall, Head of UK Government, Public Policy and Philanthropy, Twitter.

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

Katy Minshall.

Q311 **The Chair:** I am going to read the police warning; then we are off to the first question. As you know, this session is open to the public. A webcast of the session goes out live and is subsequently accessible via the parliamentary website. A verbatim transcript will be taken of your evidence and put on the parliamentary website. You will have an opportunity to make minor corrections for the purposes of clarification or accuracy. I wonder if you might introduce yourself for the record.

Katy Minshall: I am the head of UK public policy, government and philanthropy at Twitter.

Q312 **Baroness Morris of Yardley:** The first question is about authenticity. What would Twitter understand by a "fake account"?

Katy Minshall: Can I start by thanking the Committee for inviting Twitter to participate today? I wonder if you would allow me a minute or two to give a scene-setter on Twitter. It might be helpful for people who are not regular users.

The Chair: We did not have any formal evidence to look at, so that would be very helpful.

Katy Minshall: We are here today to talk about political discussion on the service, but it is important to acknowledge that people log on every day for all sorts of reasons. The most popular conversation on the service last year was actually about K-pop. We saw over 6 billion tweets about K-pop, a genre of pop music originating in South Korea.

When it comes to political discussions, we are proud that Twitter has given a public platform to many voices that, 10 or 15 years ago, lacked that opportunity, particularly the young and marginalised groups. In the 2019 general election here in the UK, we saw an increase in the conversation by 66 per cent compared with the 2017 general election. That is more people coming online and talking about politics and policies that are important to them.

The key distinguishing feature of Twitter for the purposes of this discussion is that we are the only major service to make our public conversations available for the purposes of research. Over the past decade, at least tens of thousands of researchers have accessed our public API and we are really keen to double down on that.

That is not to say that we are not acutely aware of our responsibilities here. We welcome opportunities such as this to work with experts and policymakers, and to think about what more we can do. I have two examples, if you will allow me, and then I will come to your question. The first is our announcement last year to ban political advertising globally, which I am sure we will get into. We took that decision based on the principle that political message reach should be earned and not bought.

The second is our investment in technology, which is starting to pay real dividends in the context of safety. One in two tweets that we take down for abuse we have detected ourselves, reducing the burden on victims to report these issues to us.

Let me come to your question. For fake accounts, I will talk about how we define them and then how we go about enforcing those rules. When you sign up to Twitter, we ask you for your full name and your email address or phone number. When you get on to the service, we ask you to abide by a series of rules, including that you cannot mislead others on Twitter through the use of fake accounts. The sorts of signs we take into account are the use of stolen or stock profile photos; your location — if you are saying that you are somewhere you are not; or the use of a copied bio. That rule is part of our wider policy on platform manipulation, including that you cannot mislead others on Twitter through bulk, aggressive, deceitful activity. That includes activities such as operating multiple accounts on the service to try to disrupt conversations or make something trend artificially.

The question then is how we enforce that. On the one hand, it is important that any user feels able to report their concerns on these issues to us. If they see an account on Twitter that they think is fake, they can, with a couple of clicks in-app, make us aware. However, to really get at this at scale, we have to use technology. We have been able to automate a number of processes when it comes to malicious behaviour detection. For example, an account tweeting at an extraordinarily high volume on a specific hashtag is something that a machine can detect. It might be a straightforward case of sending that account a password reset to see what is on the other end of that account. More complex cases are escalated to our site integrity team, which is charged with identifying and removing bad actors from the service.

To give you a sense of the scale, we release a transparency report twice a year where we share aggregate data on actions we have taken against our rules. Our most recent report came out towards the end of last year and we shared that we had challenged 97 million accounts in this context over the first six months of 2019.

Baroness Morris of Yardley: I shall resist responding to that, because a lot of it will be covered in follow-up questions elsewhere, but thank you for that introduction. Let us go back to the name of the account. What is the advantage of allowing people to use names that are clearly not their own or that cannot be recognised by someone else? Why do you do that? What is the advantage of doing that and what does it add to the process?

Katy Minshall: You can be pseudonymous but you cannot have a fake account. I could not say that I am Baroness Morris of Yardley.

Baroness Morris of Yardley: You cannot have a fake name.

Katy Minshall: I could not say that I am you. That would be against our rules. You can be pseudonymous. The reason is that we have seen all

sorts of really powerful use cases in countries around the world where pseudonymity has enabled people to speak out, even in the UK. Last summer we saw an account that got a lot of coverage, @thegayfootballer. Now, this purported to be a high-profile footballer who was gay and who was struggling to come out. The existence of that account sparked a wider conversation about homophobia in football.

Shortly after I joined Twitter, lots of people were engaging in a hashtag, #WhyIDidntReport. It had examples of individuals talking about why they had not reported sexual harassment and sexual assault. Being able to speak using a pseudonym enabled more voices to join that conversation and share their perspective on that issue.

Baroness Morris of Yardley: They are not all well motivated, are they? You have taken some examples, but I would still argue that it is better to do it in your name or find another way, because I am not sure what it adds to the process. A lot of abuse is done not in users' own names. If you weigh it up, without knowing the answer to the question, I suspect that far more abuse comes through pseudonyms than honourable causes such as those you have just mentioned. That cannot be the reason you allow it to be anonymous: that someone who is not quite ready to come out as gay will talk about homosexuality in football and that is the only way they can have that conversation. That cannot be the reason you have this policy. What is the advantage of it across the board?

Katy Minshall: There are a couple of questions there. If you put to one side the powerful use cases of people being able to use pseudonyms, there is the question of how you verify people's identities and whether it will work in solving the problem you have outlined.

Let us take the first question. You will be in the space of asking what ID it is appropriate to check. In the UK, you might think of a driver's licence and passport. Large swathes of the population have neither and you would not want to restrict those communities' abilities to use a service such as Twitter. Twitter operates a data minimisation approach. At this time, do people really want to be sharing more private information and data with companies such as Twitter?

Put that to one side. Will this work? There are not many use cases to draw upon here but there are a few. South Korea brought in a law in the 2000s trying to get at this exact issue where, on websites with over 100,000 users, you had to share your real identity. That law was overturned about a year later, because it was found not only to have had a negligible effect on malicious communications but to have created a database that was vulnerable to constant attempts to attack it and take that data. From my day-to-day experience at Twitter where cases of abuse come across my desk, plenty of individuals are willing to abuse not only using their full name but sharing all sorts of personal information about where they live and work.

Baroness Morris of Yardley: Why, when people register with you, do you ask people for their real name and address, if it is pointless?

Katy Minshall: There are a few things. The most important is cybersecurity. We strongly encourage two-factor authentication. That means, if I log on to Twitter from a new device, my password entry will be coupled with an email or a text from Twitter, with a code allowing me to access Twitter.

Baroness Morris of Yardley: You cannot have it both ways. You cannot say that it is really strong for that purpose but really weak in stopping people using a name that they claim is their own.

Katy Minshall: Sorry, do you mind repeating that?

Baroness Morris of Yardley: I knew I had said that badly.

The Chair: Try again.

Baroness Morris of Yardley: I will start again. In your follow-up, when I had asked why people do these abusive things in their own name, you said that it would not work, people could give false names and we could not rely on that. Then you gave various examples. You seem to be saying that, if you register in your own name with your own address and date of birth, even that is not effective. I could register as Toby Harris, pinch his ID from his pocket, and that would be that. You are saying that it is a frail system. Then I asked why you use it at all, and your second answer was that it is really robust and you ask for lots of ID. The point I was making was that you cannot have it both ways. It is either robust, in which case you can identify people online, or it is not robust. It is not a big question; it was just a clarification of what I had felt were contradictory statements.

Katy Minshall: To be clear, asking for email or phone in the context of two-factor authentication is a really strong intervention to provide cybersecurity on Twitter and services like it. On your wider question about identity online, our CEO has said that we should be sceptical of most things we see online. Even websites that purport to have a real-name policy will face the exact same challenges that I have outlined of how to identify people in the digital age.

Baroness Morris of Yardley: It is partly about the message you give out to people. If you allow them to use names that have no resemblance to their own, you are giving the message that it is okay to hide behind a pseudonym. You have made a case, which I do not agree with, for allowing it. I was not of the political era where we got as much abuse as there is now, but I would much sooner have abuse from somebody who put their name and address on the letter—it was letters in my day—than somebody who sent it with no address and with a false name on the bottom. It was just different. You might not agree with the abuse they were giving you, but you felt they were honest enough to actually state that and you could write back in whatever tone you wanted.

Are we really a better society if we make it easier for people to abuse others without saying who they are? Does that really make us a better society than saying to people, "To be honest, you could be brave and be

accountable for what you say"?

Katy Minshall: Any account that breaks our rules, whether it is a verified user or someone using a pseudonym, by abusing, harassing or engaging in hateful conduct, will be subjected to enforcement, regardless of who does it.

On your question about the experience of politicians, I have testified on this issue before and, as I said then, I do not think anyone could see the experience of politicians today and not be alarmed. For over a year now, we have had a really close partnership with the security team here and we have found that invaluable. They report abuse of Members on Twitter through a direct portal to a specific team on our side. That is coupled with a weekly phone meeting between me and the security team here. As well as unblocking issues with that process, they can give a heads-up on any upcoming debates that might be contentious, so I can tell the safety team that we could expect a high volume of reports at that time.

More recently, that has been coupled with as much training as we can offer. I was here with counterparts from Google and Facebook a couple of weeks ago running drop-in sessions for Members to talk through our approach, answer any questions and work through any specific issues they were having.

Baroness Kidron: On that last point, I think both Houses really appreciate that concentration on their particular problem. How do you deal with the broader problem of, say, women in public life, who get an incredible amount of abuse and are not necessarily Members of this House?

Katy Minshall: First, we should be working directly, as closely as possible, with expert organisations looking at these issues. The Committee on Standards in Public Life, for instance, produced a report on intimidation in public life. It made a series of recommendations, all six of which I believe we have now moved on in quite a substantial way, and we stay in touch on a regular basis to track progress. That is based on its expertise, talking about the issues of people in public life and what it wants services such as Twitter to do.

Secondly, we cannot solve problems of moderation using humans alone. We have to double down on our investment in technology. We see about 500 million tweets every day on the service and, operating at that scale, we have had to ask machines to do as much of the work as they can. That is starting to show real results. In the past year, we have increased the number of accounts we take action on by 105 per cent and that is using interventions such as machine learning.

Q313 **Lord Lucas:** Can you tell us about Twitter's experiments with labelling misinformation? Has Twitter opted for a community-led approach rather than one based on expert fact-checkers and, if so, why?

Katy Minshall: I believe you are referring to a design that was covered on news websites a month or two ago that showed, as you say, a

community-led approach to misinformation. We are exploring many approaches to misinformation. This particular one is not currently staffed and is not comprehensive of the range of approaches we are looking at. We want to work through several more iterations before considering whether to test it further.

That is not to say that we do not have a community-led approach in other ways. The first is with researchers. As I said, tens of thousands of researchers have accessed data using our public API over the past decade. About a year and a half ago, we went a step further and curated a dataset of information operation we had seen on the service over a number of years by the Russian IRA. It was about a terabyte of data that we released in late 2018 representing thousands of accounts, 10 million tweets, two million photos, GIFs and images. Since then, we have shared 22 further datasets of state-backed information operations on the service. They are available on our website right now and have been accessed all over the world by researchers thousands of times.

Why did we do that? We wanted to increase our collective understanding of these issues. In terms of the community-led approach, there are experts around the world who can bring to bear on our datasets their knowledge and their background, and ensure that we all, as a society, understand these issues better.

Lord Lucas: Are there any other approaches that you are looking at more positively?

Katy Minshall: Another key thing is how we make rules, which again is community-led. Earlier this month, we launched a new rule on synthetic and manipulated media, essentially deep fakes and shallow fakes. That rule states that you cannot deceptively share synthetic and manipulated media on Twitter with the intention of causing harm.

Where did that rule come from? It came from a public consultation. We engaged with experts and academic researchers as well as doing a public consultation. Even on Twitter, you could contribute using #TwitterPolicyFeedback and we got 6,500 responses. On another public consultation for a different rule, we had 8,000 responses. Involving our community as much as possible will mean that our policies and rules are really robust.

The Chair: Katy, does the way that, for example, President Trump uses Twitter equate to your understanding or our understanding of accountability?

Katy Minshall: Could you elaborate on your question?

The Chair: Does the way in which President Trump chooses to use Twitter conform to what we understand by democratic accountability?

Katy Minshall: I do not want to speak for your own view but there are a couple of things here. Politicians all over the world use Twitter every day to engage with their constituents and engage in policy debates.

The Chair: I was using him only as an example. I am really questioning whether very short bursts of opinion are, in any way, an acceptable alternative to being properly questioned in an open forum on the views you have. We have a danger of politicians, in particular, retreating behind Twitter tirades instead of being accountable in the way that we, certainly in this building, understand accountability to operate.

Katy Minshall: We do not see ourselves as a replacement for that. We see ourselves as complementary to those processes. In the election, we worked closely with news partners so they could live-stream things such as the debates. We had nine different live streams on election night that had over three million views. We see very direct uses of Twitter, particularly for diplomacy, where leaders across the world share their statements and reply to each other. We even saw, in the US, two politicians from different sides of the aisle finding an issue they could agree on and doing so through Twitter, but we see that discourse working alongside all the other existing mediums through which politicians engage with their constituents.

The Chair: I am struggling with the idea of Twitter being discourse but I will leave that for the moment.

Q314 **Baroness McGregor-Smith:** Could you tell us what role you believe human moderation plays in improving online experiences? How could you make the processes more transparent and consistent? I would be interested to know whether you have, or have considered, a public database of decisions you make.

Katy Minshall: Transparency is the right question. My concerns with a public database, although it is an interesting idea, are twofold. We already provide aggregate-level data about the number of reports we receive on accounts against different rules and the number of accounts we take action on. The more information we provide, the more there is a risk of inadvertently sharing private information. The more insurmountable challenge with the public database is that it could enable adversaries to build solutions designed to game our enforcement efforts.

As I said, transparency is the right question and I can talk through how Twitter thinks about transparency to our users and the general public more broadly.

Baroness McGregor-Smith: Surely, though, the right way to be transparent and to explain your values and ethics is to publish the way you moderate decisions. I watch Twitter and it is interesting how you comment on Twitter. I see it used as a vehicle for hate pretty much every day. The commentary on Twitter is not of the level that you would necessarily get in written media, because they are governed by different rules. But we do not know what your rules are. Your rules are not transparent and, as a user, I would not even know where to look at how you decide to moderate. I do not actually think you do any moderation because I cannot see it. Every time I look at Twitter, it is pretty abusive in many areas.

Katy Minshall: Our rules are on our website. We have a shortened version designed to make them as easily accessible as possible and, for each one, we have a longer version with more detail about the criteria content moderators will consider, and example tweets.

The issue you pick out is the right one. We are certainly keen to do more and think more innovatively about how we can communicate our rules and our approach. To give you an example, last month for Safer Internet Day, we made a series of videos with influencers on Twitter who had experienced abuse. They talked about their experiences and the tools they used, such as muting words in conversations and blocking accounts, to manage that every day. That had close to two million impressions with a view-through rate of 60 per cent. That is not someone clicking and drifting off; that is someone watching to the end. Those partnerships may be a more compelling way to reach people on Twitter compared with just expecting our users to go on our website and read our rules.

Baroness McGregor-Smith: If you put together a list of the things you moderated, and explained how humans moderated for you and the decisions against which you moderated, surely that would give new users to Twitter a lot more comfort about what they are opening themselves up to in this online world. Every time someone I know goes on Twitter, they are potentially subject to abuse. Although you know the rules, they do not know how you make decisions about those rules. It is a bit like walking down a dark street; you make a decision about whether you want to walk down that dark street. With Twitter, you do not know what you open yourself up to, because it is not easy to understand. Although you may have things on your website and you may argue that there is transparency, when I look at the 140 characters, including the example of the President of the United States, people are not that kind. Everyone judges kindness and humanity in a different way but, if you put this together, people would probably really admire you guys for it. Admiration is not something that I think a lot of people have at the moment.

Katy Minshall: Can I just check I understand? It would be something on our website to say what our rules are and give examples.

Baroness McGregor-Smith: It is about how you enforce them. "These are the things we have taken off. These are the things we have moderated. These are examples and this is how we enforce, humanly, our rules as an organisation".

Katy Minshall: I totally agree. We have on our website now what the rule is and the criteria that the moderators take into account.

Baroness McGregor-Smith: Do you have examples?

Katy Minshall: We have examples of hypothetical tweets. They are not real-world tweets but hypothetical ones we have come up with. We will share, where relevant, the breakdown of where we have proactively used technology to find that content and make those decisions versus where we have relied on reports from users.

Baroness McGregor-Smith: Why not do it with real tweets and say, "These ones are unacceptable so stop doing it"? Human beings listen to truth; they do not listen to vaguely moderated hypothetical examples. They think, "Right, that is the thing we cannot say". Would you at least take that away and think about how you can do it?

Katy Minshall: I am happy to take that away. I would want to check our privacy policy, but I am very happy to come back to you on that.

Q315 **Lord Harris of Haringey:** Can we go back to the question of consistency? To be quite clear, I have not checked back on the various instances where I have seen this, but I have seen people complaining about inconsistency between how their tweets and other people's tweets have been treated. What is the process of making sure you apply those criteria and your rules consistently? Would it not be easier for people to believe that you were applying them consistently if you published more information about it?

Katy Minshall: The fact that we are a public, open platform, and researchers publish information and findings on Twitter data all the time, helps get at some of these issues. The challenges on Twitter are well documented. We want to double down on research to support that and work very closely in collaboration with external experts as we find solutions.

To come back to your question on consistency, when content moderators start at Twitter, they will undertake certification exams with strict grading criteria. That will really be tested for quality and accuracy of review. Should they pass it, as their career at Twitter continues, they receive training on an ongoing basis and are spot-checked to test for quality and accuracy.

There is also a question on the product side. How do we consider safety by design in the product life cycle? On the one hand, any big product change is inherently cross-functional at Twitter. Product, engineering and design work closely with our trust and safety teams, and potentially with my team and the Twitter service team, which enforces our rules. The most important team of all is our product trust team. Product trust is a team within Twitter charged with preserving user trust and doing what it can to keep the service free of abuse and spam.

They will do that in all sorts of ways. To give you a sense, if there is a big product change, or for big products on Twitter on an ongoing basis, they have a dedicated product trust point of contact who gives timely, actionable feedback on any proposed changes. They suggest mitigations. They create new policies and systems if we are rolling out a new product. Before we even get to content moderation, we are really proud of what you might describe as a safety by design approach, where we are thinking about this from the very beginning.

Lord Harris of Haringey: I am not sure that quite answers my point about the consistency. You seem to be saying that consistency stems

from your rigorous training programme and people cross-checking. How long does somebody spend on the training programme?

Katy Minshall: I do not have that information to hand but I am happy to take that away.

Q316 **Baroness Kidron:** In this context specifically, how is Twitter finding the current misinformation about the virus? How are you doing on that and what are you learning from it?

Katy Minshall: Right now, we are not seeing significant co-ordinated attempts to spread disinformation about this topic at scale on Twitter. I spoke to the Government and DCMS last week. They also said that they are not seeing disinformation on this topic targeting the UK at this time. We will absolutely remain vigilant on that.

Baroness Kidron: Can I ask more broadly than the UK on this particular issue? It is a global virus and a global issue.

Katy Minshall: The big drive for us, at this time, is what we can do to make sure that anyone who comes to Twitter is met with credible, timely, authoritative information about the virus. We have a prompt in place so that, if you search for coronavirus, COVID-19 or a range of other trigger terms, you will be met, if you are on mobile, with a half-screen prompt directing you to the NHS website, where you will find the latest information on coronavirus. That has been in place since January. We worked with DHSC and the NHS to get that up and running. It is now in place in 64 countries across the world.

Then there is a question of what you do for users who are not looking for information on coronavirus and COVID-19 but are just logging on to Twitter. Right now, every time a user in the UK goes on their home timeline on Twitter, they will see at the top what we call a "moment". You will normally find moments in the "explore" tab—the magnifying glass on mobile. A moment is a selection of tweets that a team within Twitter has curated to depict a summary of a news story or a conversation on the service. Right now, if you click on the coronavirus one at the top of all UK users' feeds, you see the latest information, tweets from the WHO, from experts and from Governments.

Baroness Kidron: Katy, I hope I speak for everyone when I say that we all really appreciate those efforts and would love to see them mirrored in other areas. But I actually asked you whether you are finding a level of misinformation globally, co-ordinated or not, whether you are learning from it and whether you are worried by it in this particular context. I formally recognise all the good things that are going on.

Katy Minshall: As it stands, we are not seeing disinformation at scale about the coronavirus. We are absolutely remaining vigilant and working with peer companies and Governments around the world to very quickly respond to any issues that any of us see. For instance, it was in the news last week that we had worked with the NHS to take down an account purporting to be an NHS trust.

The Chair: Katy, you are a smart woman and I do not mean that in any sense patronisingly. We are in a very particular moment in time. What do you feel that the company, Twitter, and Jack Dorsey are learning about the situation we are in that would make you a better service and a more valuable public resource?

Katy Minshall: Do you mean the situation we are in with coronavirus?

The Chair: Yes.

Katy Minshall: That is a good question. I do not mean to sound like a broken record, but it is about doubling down on working with researchers. We are keen to make our data available to researchers who are looking at Twitter data in the context of contagion. Just before this committee hearing, Yale tweeted out that its researchers were using Twitter to find and source information as they think about responses to coronavirus. It has confirmed what we suspected: the importance of those partnerships.

The Chair: These are important questions for us because we are looking for good news from the digital world. Do not think that we are being negative at all. Our report will be meaningless if it is only a tirade of bad news and destruction, so I am very grateful.

Q317 **Lord Holmes of Richmond:** Good afternoon. Other platforms have stressed how their algorithms attempt to de-emphasise sensationalist or borderline content. Is this a focus for Twitter? What is Twitter doing to promote trustworthy content on its service: #thoughts?

Katy Minshall: For context, Twitter as an open, public service is in a different space here. We are not a closed, private platform where you may think more in the context of echo chambers. My colleague has a saying that the hashtag pierces the filter bubble. What do we mean by that? If you looked on Twitter in December and clicked on the general election 2019 hashtag, you would have seen a wide range of viewpoints on that hashtag. If a political leader tweets something, you see a wide range of views in the replies below. There is something there about the nature of Twitter.

Thinking about what the research says, we are proud that that is being confirmed by external experts as well. The Oxford Internet Institute did a study on the UK general election looking at the prevalence of junk news on Twitter. It found that, of the sources shared 10 times or more on the service in relation to the election, only two per cent were considered junk news and the overwhelming majority were professionally made journalistic content.

Lord Holmes of Richmond: On the general election 2019 point, it is probably worth getting that on the record. How does the service work to curate, prioritise and align what somebody would have seen if they had put in that hashtag?

Katy Minshall: In the three days before and on election day itself, we injected into the top of all UK users' home timelines a prompt that we

made in partnership with Democracy Club, a London-based organisation that you may be aware of, where you could insert your postcode to find out where your local polling station was and subsequently who you could vote for. That was clicked on 170,000 times. That shows, to your point on good news, the opportunity of digital technology as well as the challenges. That was at the top of people's home timelines.

In addition, over the course of the campaign, the curation team in Twitter that brings together the moments on the explore tab, which people have gone on to search for, curated over 120 moments of the most relevant and informative tweets covering a certain topic. Eight of those moments were focused specifically on myth-busting. Our team, which works with news partners, put together a series of live streams on election night so news partners could make the most of the service to get their content out to their audience. When you are on Twitter and you are looking for information about the general election, there was plenty of high-quality, informative, credible content.

Lord Holmes of Richmond: In that process, what is the determinant and who determines relevance?

Katy Minshall: I can explain how our algorithm works in general, which would apply to the election as well. On the Twitter home timeline, you may have the ranking algorithm turned on. Historically, there was no algorithm on the Twitter home timeline. You would just see, in reverse chronological order, tweets from accounts you followed. That was the case for most of Twitter's existence. In 2016, we brought in an algorithm, after extensive user feedback that seeing tweets in that way was not helpful. People did not want to have to log on every five minutes to keep on top of Twitter. They wanted to be free to log on however often they wanted and see the tweets most relevant to them.

The key thing is that we give all users the option to just turn the algorithm off. If you are on mobile, in the top-right corner there is a sparkle icon where you can turn the algorithm off and once again see tweets in reverse chronological order from accounts that you follow.

Lord Holmes of Richmond: That is very helpful. Thank you.

Q318 **Lord Lucas:** I have done that because I found that your algorithm was hiding a lot of my right-wing friends. There is a very wide spectrum of people I follow on Twitter and it had become very politically emphasised. Is there more you could do on Twitter to make it a friendlier place for debate? There are a lot of sentiment analysers people use on Twitter. Could you use that yourself so that you are emphasising kinder conversations? Could you do it yourself so that someone who has started a thread can tag a reply lower down, say, "This is good; this is authoritative" and bring it back towards the top of the thread?

Katy Minshall: These are exactly the kinds of questions we think about every day. Our aim is for the public conversation to be as healthy as possible. What we mean by that is healthy debate, conversation and

critical thinking. The converse of that is abuse, spam and platform manipulation. In terms of how we get there, yes, we absolutely think about what we could do to bring to the surface the most engaging, interesting responses.

In the way our algorithm works, you do not necessarily see replies in chronological order. If you tweet something out and you respond as the tweet author to one of the replies, that is at the top. If someone you follow replies to your tweet, again, that is brought to the top because, as you point out, the chances are that that will be more relevant and interesting to people who want to keep track of the conversation you are having with your audience.

Q319 **Lord Lipsey:** Facebook put out a white paper in February that focused on procedures and processes as a way of dealing with the various problems we have raised. Your response seemed to hint that you took a slightly wider view of things. Is that right?

Katy Minshall: Our CEO has said that we absolutely welcome smart regulation. The key thing is that we are not waiting for regulation. I work closely with colleagues in DCMS and other government departments all the time as they think really carefully about the Online Harms White Paper. I think this was first proposed in 2017 and we are really keen to move on the concerns people have as regulation has developed concurrently.

To your question on systems and processes, Facebook's white paper made a compelling point. If you just focus on the outcomes, such as the number of reports you get about X or the percentage you act on, there is a very real risk of creating perverse incentives. It might encourage a company, for example, to narrowly define its rules. It might encourage a company to make it harder to make reports because it is not prepared to deal with the big volume. It might encourage a company to focus overwhelmingly on responding to reports at the expense of proactive work and investing in technology that could help get at the problem at scale.

It is worth regulators considering a more holistic picture, looking at systems and processes, which seems to be the view of the online harms White Paper here anyway.

Lord Lipsey: As a company, would you welcome greater regulation from regulatory bodies on the grounds that it would create a level playing field or are you fearful of the impact on free speech for which Twitter is such an ally?

Katy Minshall: We welcome smart regulation. In our view, freedom of expression means that all users feel safe expressing their unique point of view. If collaborations with safety experts, with governments and with whoever help us get there, that is something to welcome.

The Chair: We took evidence on this, as you know, from the ICO and from Ofcom. We discussed the process of regulation and, indeed, we

discussed penalties. I could not make head or tail of where Facebook sits on appropriate penalties. Where do you sit? By that I mean, with anonymous use of Twitter, you know who the people are. If somebody has been continuously abusive towards an MP or somebody else, what form of penalty would you regard as acceptable and who would impose the penalty?

Katy Minshall: I saw some of the testimony this morning. On Twitter, we can suspend someone from the service. We have to work very closely with lawmakers and law enforcement if they want to take next steps. With malicious communications, I believe only three per cent of cases result in a charge. There is a need for all sectors to work together and think about what appropriate penalties are in place for behaviours that are illegal both online and offline.

The Chair: The ICO made it clear that, now she has the power to charge up to four per cent of revenues, her position vis-à-vis the people she is regulating is dramatically different from when she had a ceiling of £500,000. Are you in favour of higher penalties to get better regulation and, indeed, more compliance with regulation? Do you feel that freedom of speech then gets squeezed out and there is a problem at that end of the spectrum?

Katy Minshall: Our global data protection officer has met with the ICO every six months for a long time now, since before this change occurred. The detail of regulation really matters before you even think about enforcement action. Clarity would be really welcome when a company such as Twitter adapts to a new framework. Do you have a good point of contact with law enforcement? Do you make your rules publicly available? Do you do proactive work to convey those rules to your users? Those sorts of systemic process requirements are more straightforward for companies to know what they need to do. When you get into the detail of trying to define something as complicated as misinformation, it puts a company in a more challenging place in understanding what it would need to do to comply with that regulation and avoid the risk of fines or equivalent measures.

The Chair: At the risk of being a bit boring, I am fascinated by the issue of regulation and particularly the history of the media. I was quite surprised that one reason freedom of the press was established in the middle of the 18th century was that a lot of anonymous pamphlets were published but the person who took responsibility was the printer, and the printer's name had to be on the pamphlet. A lot of printers were prosecuted but the courts continually found the printer not guilty. It was the courts deciding in favour of freedom of expression that created the environment we have in the media today. I just wonder what the parallel would be, frankly, in the online world.

Katy Minshall: This is an interesting question. I would not want to say anything with certainty without reading more into the scenario you outline. I guess you are getting at this question. Could there be an opportunity with regulation for people to have more trust in services such

as Twitter, because a regulator is saying, "You are doing what you said"? There is value to that but, as we are trying to be transparent with Ofcom, which the Government are minded to appoint as the regulator, that does not solve the problem of transparency with users directly. Academic research and proactivity in your communication to users are important.

Q320 **Baroness Kidron:** In a way, do we look at this with a blunt instrument: regulation or no regulation? Is there not a sense in which the culture of Twitter should be really transparent and upheld, so what you say you do in your terms and your community rules is done, and the job of the regulator is to come in and ensure you have done what you have said you have done? There is then a two-step process. In this conversation, I hear the argument about freedom of expression, but you are a private company. You have your rules, standards, corporate values and so on. Even your algorithm seems to be rather to the left, we find out. That is also fine, so long as we know what you are and you are what you say you are.

In a way, the job of regulator is slightly different from the one it is always cast in, which is saying, "Okay, you have said what you are and you are doing what you said, so you get a big tick". There may be things at the edges about illegal content and so on, but in my experience, both as a company and through the Internet Association, techUK and various other forums, there is always a pushback about regulation and not a huge amount of transparency. I appreciate all the positive things you have said but, on this point, is that a healthier way for us to look at it as a committee: we hold you to account for what you say and give the regulator teeth and possibly a fist to make sure you do what you say?

Katy Minshall: This is the opportunity with the approach you outline. If we had fixed in law what we thought of as priorities for the internet space in 2017, the chances are that there is a whole host of work that that would not have captured. The risk of the regulator, Ofcom, or the Bill itself trying to fix specific codes or requirements on a whole multitude of issues is that it will not be as future-proofed as it could be. If you adopt the approach you outline, trying to get at whether companies are doing what they say they are doing, companies might be better able to respond to emerging issues more rapidly.

Baroness Kidron: For clarity, there always has to be a floor if not a ceiling.

Katy Minshall: Yes.

Q321 **Lord German:** Can I turn to what you do allow the users of Twitter to do, which is to hide the replies to their tweets? Can you tell us how many people, either globally or in the United Kingdom, have taken up that opportunity and are using it?

Katy Minshall: To give some context, we want everyone on Twitter to feel safe expressing their viewpoint. To get there, we took the view that we had to change how conversations were working. The issue, as we have talked about, was that repliers were detracting from the tone or

topic that a tweet author wanted to have with their audience. Last year we took the decision to trial the ability to turn off replies in Japan, Canada and then the US. Just a few months ago, we rolled that out globally to all users. I do not have the results of the global rollout yet, but I have the results of the trial if you would like me to read them out.

Lord German: Yes, please.

Katy Minshall: People mostly hide replies that they think are irrelevant, off-topic or annoying. The option is a new way to shut out noise. Of the people who hide replies, 85 per cent are not using block or mute. People were curious to see how public figures like those in politics and journalism would use this update. So far, they are not hiding replies very often. In Canada, 27 per cent of people who had their tweets hidden said they would reconsider how they interact with others in the future. They also thought it was a helpful way to manage what they saw, similar to muted keywords. We learned that you may want to take further action after you hide a reply so we now check to see if you also want to block the replier. Some people mentioned that they did not want to hide replies due to fear of retaliation as the icon remains visible. We will continue to get feedback on this. These are early findings and we look forward to continued learning as the feature is used by more people.

Lord German: I did not quite get from that the percentage of users in Canada and Japan who took up that option.

Katy Minshall: I do not have those figures for you today but, after watching the hearing this morning, I have contacted colleagues in the US to see if there is further data we can provide for you.

Lord German: Thank you very much for that. The other issue related to this is how easy it is. I was interested that you said politicians in Canada used this more often than anybody else. Of course, politicians are quite robust people, although we do not like being trolled and abused. The people who have an opinion, lay it out and then get abused find it most difficult to sort this out. You have a choice, you say, to block or hide; it can be reawakened. Is that all user-friendly? If the people using it are mostly the politicians and the sort of people who are expecting to get that response, clearly it is not good enough for the people who feel most hurt by it.

Katy Minshall: To clarify, we saw that public figures, such as those in politics and journalism, were not hiding replies very often. To your question about what next and what more we do with this, we announced in January that, partly as a result of this, we would now look to experiment with the ability to turn off replies or to specify who could reply to your tweet. At the moment, if you tweet, anyone can reply to that. We are going to trial conversation controls. You can keep it as it is; you can have it so that only people you follow can reply to that tweet; you can, for example, invite the Committee here to reply to your tweet; or you can turn off replies completely. The thinking is that putting those controls in place or making them available might encourage people to feel

more confident about tweeting, because they are less concerned about the barrage of replies they may get below.

Lord German: Can I go back to a question you spoke about right at the beginning? I understand that hiding a response can automatically happen, or there is a tool for making it happen, where you do not have an email address, phone number or profile picture. From the conversation at the beginning, I thought that you had to have their email address or phone number, to know that they were a real person behind whatever was said. This is not really a user device, because the only one that knows whether you have a registered email address or phone number, or are not changing a profile picture, is Twitter. Is Twitter doing this on request or do you do it automatically?

Katy Minshall: This is the ability to hide replies.

Lord German: Yes. It is also about users restricting the notifications they receive so they are not informed when they are mentioned by someone who has not had their account authenticated. I thought you had said that everybody who had an account had to have it authenticated.

Katy Minshall: You are absolutely right. You can control what notifications you get from Twitter. This is particularly useful to politicians and others in public life who get high volumes of notifications. As you say, you have the option to get notifications only from people who have verified an email address. You can also turn off notifications for everyone except those you follow.

By the way, tying back to your earlier comments, we want to do as much as we can to make people aware of the tools. A good thing, which may be a challenge as well, is that on Twitter you can mute, block and hide replies, limit notifications and, in months to come, turn off conversations. That is a lot for users to take in. We are going to look for really good opportunities to clearly articulate how you can use these safety tools and in what context.

It is important to mention that we are looking at whether external developers can make use of the "hide replies" API so you can automatically control what replies are filtered. For example, a safety organisation might create a plug-in where it would automatically hide replies to your tweets that included certain words. That is a really blunt example, but it gives people more control and opportunity to shape their experience on Twitter.

Lord German: I just want to get at the very important point that I was repeating from an answer you gave earlier. You said that all accounts had to be authenticated. Here is this device with which users can switch off any unauthenticated account. Why would they have an account if you had not authenticated it?

Katy Minshall: Not all bots are bad; that is one way of putting it. We will ask for an email address or a phone number, but you can operate

multiple accounts. You can have a bot on Twitter without necessarily a specific human on the other end. To give you an example, there is a bot right now that reminds people to wash their hands and goes off every so often. To give you another example, last year there were lots of petitions related to Brexit. The petitions website crashed a few times because so many people were going on it to see how many people had signed up. There was a bot that would, every 15 minutes, give the updated number of signatures, trying to prevent that website from crashing by giving people the information they needed.

Yes, we ask for an email address and a phone number, not least for the purposes of cybersecurity, but historically people have created accounts on Twitter that are bots, with no human on the other end, for good reasons.

Lord German: You allow bots.

Katy Minshall: Yes. We do not have a rule against bots. We have a rule against automated accounts and multiple accounts trying to spam a conversation or engage in a malicious way on Twitter.

Lord German: Someone will have to explain to me the difference between a bot and an automated account, but there we are.

Lord Lucas: Looking at how we should keep tabs on all these changes going on and their effects, would you be content if Ofcom had a duty to comment on the way in which you ran your site and whether it was supportive of or antithetical to democratic engagement of citizens? In order words, we give Ofcom a general duty to keep an eye on what you are doing and say, "Hang on; the way you do that is causing problems. Why not do this, because it would make things better?"

Katy Minshall: That is the right question to ask. We should all be talking about how we are engaging with democratic discussions. It is an interesting idea. I did not watch the Ofcom testimony. If they had some thoughts, it would be interesting to read them. If you try to specifically define misinformation, which therefore translates to what services such as Twitter do or do not take down, and what platforms should or should not do, I can imagine some issues in the context of democracy. It might be difficult, but your question is the right one. Yes, every service should be thinking about how it reflects political discussions and engages in democratic processes.

The Chair: I think you might find yourself quoted in the report.

Q322 **Baroness Kidron:** Can I ask you about the new mechanics of conversation hiding and so on? Are you not concerned about what we have seen on other platforms where conversations get private, people try to monetise that and powerful people in the space ask for money to get into the private conversation? That is one version. Another version is even less savoury about grooming and so on. I am interested, and I may have misunderstood what you said, but you seem to suggest that there

are a whole number of things that could mean you go out of sight in some way. The answer may be no, but I am interested to know whether those considerations have been in mind.

Katy Minshall: Much like our political advertising ban, we very much stand behind freedom of expression and freedom of speech, but freedom of reach is different. To be totally clear, what I have talked about is in no way a change to Twitter's public, open nature. Even if you were to hide my reply, that does not prevent me tweeting that you have hidden my reply. It does not prevent me sharing that viewpoint.

Baroness Kidron: Nothing you have described is actually to the side of the public platform ultimately.

Katy Minshall: Correct, yes.

Q323 **Lord Harris of Haringey:** You reminded us earlier that political message reach should be earned, not bought, so I am going to ask about that. You have banned political advertising. I understand Mike Bloomberg, in his brief and ineffective presidential campaign, paid influencers to promote his campaign and paid regular members of the public to tweet positively about him. Are you going to stop that and, if so, how?

Katy Minshall: The political ban was informed by three principles. As you have just said, political message reach should be earned and not bought. In addition, advertising should not be used to drive political outcomes. The interaction between civic conversations and microtargeting is not yet sufficiently understood and poses new risks, so we should adopt a cautious approach. We already ask that any user posting an advertisement as an organic piece of content discloses its commercial nature to users. We have best practice on our website, such as saying "#ad", and we will continually evolve our rules and processes as the nature of content changes on Twitter.

The key thing here is that, if a political party makes an arrangement with a high-profile individual or celebrity to post something on Twitter, it still has to earn that reach. It cannot buy the reach on our service.

Lord Harris of Haringey: If it pays the celebrity, who has a high reach because they tweet about football or whatever else, is that not the same?

Katy Minshall: The question is the right one. The challenge is that we, Twitter, would not have visibility of an arrangement made between a political party and an influencer or high-profile individual. That political party in the UK should be disclosing its digital spend to the Electoral Commission. This Committee is talking to Twitter, Facebook and YouTube but in the 2017 general election, not just the most recent one, we saw political parties creating their own apps. There are now over two million apps on the Android app store. The way people engage with democratic life through technology is not fixed. Platform by platform, service by service, may be the wrong way of looking at it, when there is already an Electoral Commission that can make these decisions and enforce them in that way.

Lord Harris of Haringey: What I am trying to get at is whether there are ways to circumvent your rules. We have seen the phenomenon of Twitter storms to get particular political messages across. Those are not paid for but may use a network of activists or whatever else. Would you see those things falling into the category of political advertising?

Katy Minshall: Sorry, can I ask what you mean by Twitter storms?

Lord Harris of Haringey: I was rather hoping you would know because, when I hear there is going to be a Twitter storm about a particular issue, I try to keep away from it. As I understand it, hundreds or thousands, depending on the organisation concerned, deliberately tweet on the same topic at a specified time, so they get a trend and the trending sign appears. This is then seen as an effective way of campaigning. That can be orchestrated either by the sheer enthusiasm of a political party's supporters or by paying people to do it for you.

Katy Minshall: It is already against our rules to artificially amplify a conversation. Political party or not, you cannot create a load of accounts and ask them to all tweet the same thing at the same time, and like each other's tweets, to try to get something to trend artificially.

Lord Harris of Haringey: How do you stop it?

Katy Minshall: To go back to my earlier response to Baroness Morris, it is a mixture of people reporting it to us and our proactive technology. A computer can detect extremely high-volume tweeting on the same hashtag and intervene.

Lord Harris of Haringey: Hang on. COVID-19 is no doubt a very popular hashtag at the moment and it has suddenly peaked. You do not pick that up but you would pick up if somebody was tweeting something about universal credit.

Katy Minshall: Whether it is about COVID-19, universal credit or any topic, if someone or a group of people creates a series of fake accounts to try to orchestrate or amplify a conversation, that is against our rules.

Lord Harris of Haringey: Does it get taken down and just disappear?

Katy Minshall: There are all sorts of ways we could intervene. We could reduce the visibility of the trend. We could find and remove the accounts, and suspend them from the service. Yes, we would intervene in multiple ways.

Lord Harris of Haringey: How quickly could you do that?

Katy Minshall: It varies. We want to be as quick as possible and stop things before they reach the point of trending. In other cases, we would ask people to report that to us.

Q324 **Baroness Kidron:** I am not sure you have answered the first half of my question, but I want to give you the opportunity to say anything you feel

you have not said, which is really about how you come to design and roll out new products and the forces that might make you choose to do so.

Katy Minshall: Our aim is for the conversation to be useful, interesting, timely and informative. The products we roll out, the products we change and the products we deprecate are trying to serve that purpose. As well as the work of the product trust team, which I outlined, it is worth talking about our prototype app. This is something we launched last year and invited ordinary Twitter users to join, where we will test new designs or products directly with users, to get feedback from them and to see how a particular change might roll out in reality before we consider rolling it out on the service writ large. As well as the internal safety by design approach, we have an external test bed to look at these new innovations.

Baroness Kidron: I want to ask you about safety by design. I understand that methodology, but so many problems in this area have been dealt with retrospectively. Could we look at the development of Twitter and say, "In August 2017, it changed from looking over its shoulder to looking forward in terms of its product"? Has there been a change of attitude and what happened to cause that?

Katy Minshall: I have a timeline of safety changes we have made over the past few years, which I will provide to you after this hearing. No one at Twitter says that we have everything perfect—far from it. There is a reason why, if you go on our blog, the title right now talks about our safety and it says, "Progress and more to do". Because we are a public open platform and there is so much research, the challenges are well documented on Twitter. I am sure that, in the way that we have already changed a lot over the past few years, particularly using technology more and more, we will continue to change and adapt again. We are starting to see really significant results that give us confidence that our approach is paying dividends.

A few months ago, we announced that we had seen a 27 per cent drop in bystander reports. If I was abusing Baroness Morris and Baroness Kidron reported it, that would be a bystander report versus Baroness Morris reporting it herself, which would be first-person. That drop in reports suggests that people are seeing less abuse on Twitter than they used to. That gives us confidence that the investments we have made and the technology we are putting in place is having an impact. Like I said, no one at Twitter would say we are perfect. No one would say we do not have work to do.

Baroness Kidron: I would be very interested in seeing your timeline. If you were willing, I would be interested in your view about what preceded the change, whether it was regulatory, whether it was reputational, whether there was a particular incident and so on. That would be very interesting and you may indeed star in our "good news" chapter.

I want to ask about the business model. I do not want to ask just about Twitter. If you think about organising all the world's information, connecting all the world's people or creating a fantastic public

conversation, these are all things that everyone in this room would like to see happen. But it has turned into a bit of a war of attention, growth and user numbers in order to advertise. I wonder whether you have anything to say about the culture wars, if you like, between your mission on paper and the need for growth because of the business model. Do you consider that there is another business model available for Twitter?

Katy Minshall: Principally, our business model relies on people coming on to Twitter every day. That is how services such as Twitter succeed. People are not going to come on if they do not feel safe. They are not going to come on if they are not seeing useful, relevant, high-quality information and content. Our business model and our priorities as a company of safety and health are absolutely aligned.

The Chair: You are suggesting that people do not come on to Twitter in order to shout, abuse and insult but that they come on to listen.

Katy Minshall: People come on to Twitter for all sorts of reasons. We define 40 per cent of users as listeners, in that they do not tweet, but they are there, interested in the news, sports and entertainment. I talked about K-pop. We published research last year looking at the UK specifically and the communities that have formed. There are the big ones that you might expect, such as gaming, sports and football, but communities have also formed around cheese, wine, gardening and bird-watching. We are here today to talk about political Twitter, but there are so many different forms of Twitter that go far beyond talking about politics or Brexit every single day.

The Chair: Is that not the reason why we should, in theory, be helpful to you? We are talking about the potential reputational damage that then disrupts or upsets all the good things you could be doing. Because of exactly what is happening on social media, the reputational damage argument dominates in exactly the same way that, unfortunately, on too much social media, the negative platform manipulation dominates. It may sound strange but we are endeavouring to be helpful to you, because we believe that, as a sector, you have a reputational issue. You at Twitter probably have it less than anybody else.

Lord Harris of Haringey: While Baroness Kidron was asking her question, I typed in the search term on Twitter "#TwitterStorm". What immediately came up was a tweet from @Artists4Assange yesterday saying, "Why should #JulianAssange be urgently released from British custody? Please answer in your own words & retweet! If we get 100 retweets, @Artists4Assange will launch another #TwitterStorm calling for Assange's immediate release". Now, is that a manipulation and, if it is a manipulation, is it something you pick up? Obviously you did not pick it up and, as far as I can see, it got 96 retweets, so it did not meet its own criteria. The point is that, if you have a community of people, you would use something like that to try to amplify it. Are you saying that you would or you could stop it?

Katy Minshall: I do not want to give you any incorrect information, so let me take that example away.

Lord Harris of Haringey: I suspect that that example is irrelevant. It is clearly a tactic that people are aware of. Similarly, the next one is about finding a lost dog, #FindBertieLakeland. There is a picture of a very fetching dog. All I am saying is that there seem to be people on Twitter who see this as a tactic to maximise their message. Finding the dog may be a very beneficial one. With Julian Assange, it depends on what you think about him. It seems to be a manipulation that people are well aware of. Can you track and deal with this, if you think it is unhelpful, abusive or anything else?

Katy Minshall: Yes. To your point, there is a difference between someone retweeting a missing pet or a missing person, encouraging people to share to get as much reach and exposure as possible, and me, with my friends, saying, "Let us create a load of fake accounts and get this particular hashtag to trend", which would be completely artificial in nature. There is a difference between that and saying, "I am watching a film. Retweet and tell us what your favourite film is". That would be an authentic engagement versus disrupting a conversation on particular terms in a very non-genuine and artificial way.

I am sorry that that is not a very black and white answer, but that speaks to the fact that it is not a black and white tool. Bad actors game it whereas others use it every day for good.

Baroness Kidron: We often get the cast of bad actors. I absolutely agree that they are out there and it is troublesome to deal with them. The Committee struggles with the more systemic issues. Where is the system itself creating a problem or unfairness? I want to quote from the developer who invented the retweet. He said it was like handing a loaded weapon to a four year-old, due to the way the feature prioritises sharing content at speed rather than encouraging users to read or think about the content they are sharing. He is making the point there, from an internal perspective—and, as was pointed out, the more traffic there is, the more Twitter is winning—that there is something intrinsic in the conversation that is destabilising to productive conversation. That is slightly different from whether people who like cheese are tweeting about gorgeous cheese.

Katy Minshall: Our CEO has been pretty upfront about this. He has said he is interested in rethinking the fundamentals of the service. The fact that we are considering rolling out the ability to just turn off replies to tweets speaks to that. Yes, we absolutely think about the fundamental parts of Twitter. Again, this is where we have to work with researchers. We have a partnership with UC Berkeley right now, looking at the interaction between machine learning and social systems. What impact does it have on people when you have a certain algorithm in place? Does it cause them to change their behaviour?

Q325 **The Chair:** That is why I challenged you on the use of the word

"discourse". It is not discourse at the moment. As long as we know it is not discourse, we are not likely to be quite as confused as we sometimes are. Tweeting is much closer to a football crowd chanting, but that is a personal view.

In what ways are you working to improve academic access to Twitter data?

Katy Minshall: I have already talked about our API and the information operations datasets we now have on our website. The step change we are really keen to work on now, rather than just making data available, is to work directly and closely in research partnerships. We have a dedicated team at Twitter working with academic researchers. In January, we launched a new hub for academic researchers so they could quite quickly find out what information they could get from Twitter and how to get there.

Last week, as a result of feedback from our engagement with the academic community, we made some changes to our developer policy. They said that it would be helpful if they could share an unlimited number of tweet IDs or account IDs with their peers for the purposes of peer review, and our policy was getting in the way of that so we have added that clarification.

As a good example of the direction of travel for us, which is on our website right now, last week we announced a new partnership with the New Zealand National Centre for Peace and Conflict Studies. It did some work looking at the aftermath of the Christchurch massacre and the way Twitter was used, overwhelmingly to share grief and outrage at the atrocity. Its research with us is going to focus on how social media can be used to promote inclusion and tolerance, and fight against exclusion and intolerance.

We are keen to double down on our partnerships with researchers. Thinking back to our discussion about regulation and the White Paper, what would be great, before the next iteration of the online harms Bill or White Paper, is detail on how companies could be incentivised to work more closely with the research community.

The Chair: Your company should be very proud of you. Thank you very much.

Katy Minshall: Thank you.