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

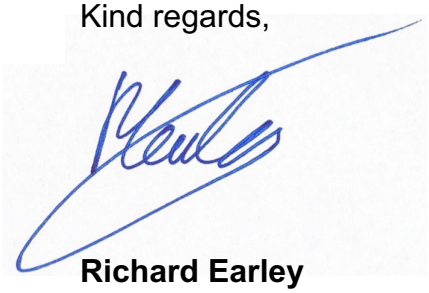
These policies have been developed in consultation with outside experts. They do not use the terms “non-physical, indirect, or vicarious harms”, but we continue to talk with these experts about the ways different content on our platform might lead to harms, and how we should describe and take action 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