[COR0178]

Written evidence submitted by Facebook (COR0178)

Facebook welcomes the opportunity to respond to the Home Affairs Select Committee's call for evidence into COVID preparedness and specifically online harms during this period.

We understand that in particular the Committee is seeking evidence on the nature, prevalence and scale of online harms during the Covid-19 period; steps that could be taken to mitigate these concerns; and the adequacy of the Government's Online Harms proposals to address issues arising from the pandemic, as well as issues previously identified.

We set out below an overview of the evidence we have of the nature and scale of online harms during the COVID-19 period. We have provided details of the steps we take to address these harms, including those that we have introduced specifically to deal with new incidences arising during the pandemic. Finally, we have laid out where we believe the Government's Online Harms proposals would help to address these issues, and areas we believe may require further development.

1. Introduction

Facebook was built to help people stay connected. Our mission is to give people the power to build community and bring the world closer together. We're committed to building technologies that enable the best of what people can do together. Our products empower more than 2 billion people around the world to keep in touch, share ideas, offer support and make a difference. Over $2 billion has been raised by our community to support the causes they care about, over 140 million businesses use our apps to connect with customers and grow, over 100 billion messages are shared every day to help people stay close even when they are far apart and over 1 billion stories are shared every day help people express themselves and connect.

The safety of our users is our most important responsibility, and we take the presence of harmful content on our platform extremely seriously, and this is why we have developed a range of tools and approaches for tackling it, including recent measures to deal with the exacerbating factors of the COVID-19 crisis, which we will outline in this response.

We work closely with trade bodies and partners, supporting industry-led initiatives to share knowledge, build consensus and work towards making our online platform as safe as possible. We have said for some time that we would welcome a more active role for Government as these are complex issues and we believe that well designed regulation would be a benefit to us all, reflecting this view in our response to the UK Online Harms whitepaper.

1.1 Our Community Standards

Facebook has had rules about what content is and is not allowed on our platform for well over a decade, which we call our Community Standards and we develop and iterate these rules with a wide range of experts and partners.

We have over 35,000 people working in safety and security, with around half of these being content reviewers responsible for actioning reports made to us and enforcing our rules. To be transparent about the progress we are making against harmful content we issue a quarterly transparency report, which includes a Community Standards Enforcement Report that shares metrics on how we are doing at preventing and taking action on content that goes against our Community Standards. These actions may include removing content or in other instances covering content with a warning screen.

We prohibit content relating to child nudity or sexual exploitation of children (CSE), terrorism and organised hate, and content detailing violence or criminal acts. We also prohibit content that is not illegal, but may be harmful for our users. This includes misinformation and disinformation, bullying and harassment, adult nudity and sexual content, fake accounts, graphic content, and content relating to suicide or self-injury.

We prioritise 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. At this time, it also includes harmful content related to COVID-19.

ur latest Community Standards Enforcement Report covered the period from January 2020 to March 2020. It showed that 99% of the violent and graphic content that we took action on was found by our moderators or algorithms before any users reported it. We also raised the proportion of terrorist content that we actioned before user reports from 99.1% to 99.3%. The full report details our actions across a number of different types of content.

1.2 Identifying and removing harmful content

For much of Facebook's history we relied on manual removal of content that was reported to us by users. This approach has enabled us to remove a lot of harmful content, but it meant we removed less harmful content before people saw it. Moving from reactive to proactive handling of content at scale has only started to become possible in the last few years because of advances in artificial intelligence -- and because of the multi-billion dollar annual investments we have made into this technology.

Today we use computers for what they're good at -- making basic judgements on large amounts of content quickly -- and we rely on people for making more complex and nuanced judgements that require deeper expertise. Some categories of harmful content are easier for AI to identify, and in others it will take more work to develop the technology. For example, visual problems, like identifying nudity, are often easier for AI than nuanced linguistic challenges, like hate speech. Our systems already proactively identify 93.8% of the nudity we take down, up from just close to zero a few years ago. While proactive detection of hate speech remains lower, we are making progress, our proactive rate has climbed to 88.5%, from 68% at the start of 2019. We anticipate these figures improving further with advances in technology.

It's important to note that proactive enforcement doesn't change any of the policies around what content should stay up and what should come down. That is still determined by our

Community Standards. Proactive enforcement simply helps us remove more harmful content, faster.

2. The nature, prevalence and scale of online harms during COVID-19

2.1 Summary of recent Global trends

In Q1 of 2020 we actioned 9.6 million pieces of hate speech. We are now able to proactively identify a much larger proportion of this, removing 88.8% of it before a user reported it, compared with 68.4% a year ago.

In Q1 of 2020, we disabled 1.7 billion fake accounts. Almost all of these (99.7%) were found and flagged before our users reported them.

We removed 8.6 million pieces of content relating to child nudity and exploitation violations; we took action on 25.5 million pieces of content relating to violent and graphic content violations and we removed 1.7 million pieces of suicide or self-injury pieces of content.

2.2 Misinformation relating to COVID-19

The most immediate online harm during the COVID-19 crisis has been misinformation. In the context of coronavirus, we define misinformation as including:

- **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 ofCOVID-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, claims that essential services are now or soon to become unavailable, unless the appropriate governmental authority has publicly confirmed thatinformation.

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. When people saw those warning labels, 95% of the time they did not click to view the original content.

3. Steps to mitigate online harms during COVID-19

Facebook is working proactively with Government, medical experts and civil society to help defeat the Coronavirus pandemic.

Since the outbreak began we’ve been working night and day - globally and here in the UK - across four key areas:

1. Delivering public health messages to the public
2. Tackling emerging harms (such as misinformation)
3. Investing in tailored solutions across each of our services
4. Working with academics and government to stop the pandemic

The first three of these areas are the most relevant to tackling online harms during the COVID-19 period. The vast majority of people are using our services to keep in touch with friends and family, help one another and access NHS and other Government information. But the crisis is far from over and there is more we can do as we move into the gradual easing of the lockdown restrictions.

3.1 As a technology company with millions of UK users, delivering public health messages to the public is one of the most valuable services we can offer.

We have worked to ensure that Government and medical community messages and guidance are delivered to our users as people log onto our platforms to speak to friends and family and seek medical information.

Since the start of the year, Facebook has been connecting its users with the most authoritative, up to date information about the coronavirus from trusted sources including the NHS. Anyone who searches for information about the virus on Facebook or Instagram receives a pop-up that directs them to the latest GOV.uk guidance, either directly or via our dedicated Coronavirus Information Centre on Facebook. This Centre brings together the latest posts from Government and health authorities, news from trusted sources, verified statistics and figures, and internationally recognised guidance.

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.

Globally, we launched an external hub for Governments responding to COVID-19 that they can use to maximise the reach of their public messaging. This includes tips and tools for public ervice announcements on Facebook platforms; advice on emergency response and reparedness for online communities; guidance on how to use Facebook groups to connect with constituents; and a guide to our ad policies to avoid disapprovals of important announcements.

In the UK we have taken a number of specific steps:

- We include NHS, Government and PHE information in the Coronavirus Information Centre which is promoted at the top of people's News Feeds;
- We launched a Government automated WhatsApp service as a source of official information, and published a guide for public bodies that want to communicate with people about COVID-19 over WhatsApp, including how to create default answers to frequently asked questions. This service has sent more than 1 million messages;
- We provided Public Health England (PHE) with free advertising credits to enable them to supplement their other campaigns by also reaching out to the 43 million adult Facebook users in the UK with crucial coronavirus messages;
- We have written to all Members of Parliament outlining how they can most effectively share official Government guidance with their constituents utilising their Facebook or Instagram accounts. We plan to write again with updated advice as we continue to adapt our tools during the pandemic. This will be a regular communication and we have made clear to MPs that there is an open channel communication to our team at all teams.

3.2 As mentioned, one of the most prevalent online harms during the COVID-19 period is misinformation about the virus, including its provenance, its impact, how to treat it, and other connected theories such as its relationship to 5G.

Misinformation is a complex and evolving problem, not least during the COVID-19 period. To address it we follow a three-part framework:

- **Remove** - We remove content that violates our Community Standards.
- **Reduce** - We work with more than 60 partners around the world, factchecking content in more than 50 languages. When content is rated false, we dramatically reduce its distribution so significantly fewer people see it.
- **Inform** - We also believe it's important to inform users when they encounter misinformation, so they can decide for themselves what to read, trust, and share. If content is rated false by a fact-checker, people who see it, try to share it, or already have, will see warnings alerting them that it's false.

We have taken a number of specific actions to tackle misinformation stemming from COVID-19. 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 January, we have applied this policy to misinformation about COVID-19 to remove posts that make false claims about cures, treatments, the availability of essential services or the location and severity of the outbreak.

For claims that don’t directly result in physical harm, like conspiracy theories about the origin of the virus, we continue to work with our network of over 60 fact-checking partners covering over 50 languages to debunk the claims. In the UK, we are working with Full Fact and Fact Check NI who are reviewing content and debunking false claims and we recently onboarded Reuters as an additional Fact Checking partner. We are also partnering with the International Fact-Checking Network (IFCN) to launch a $1 million grant programme to help fact checkers scale up their work at this time.

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.

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. We also send similar information to users who had previously shared the content that has been debunked, via a notification.

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 have also provided the Cabinet Office with a bespoke dashboard from our partners at Crowdtangle that allows them to gain insights into trending public COVID content from across Facebook and Instagram.

Following recent guidance from health authorities and governments, under our existing policies against harmful misinformation, we are removing false claims that link 5G to COVID-19, such as claiming that 5G is the cause of COVID-19, the symptoms of COVID-19 are really a result of 5G, and that government social distancing order are just a ploy to install 5G infrastructure. In addition to our misinformation policies, we have already been removing content which encourages attacks on cellular towers or 5G masts. These policies apply globally.

We are using our automated detection systems to help limit the spread and remove this kind of content as well as working with our global network of independent fact-checking partners to fact-check general 5G hoaxes.

3.3 Unfortunately, given the scale of the crisis, a number of other on and offline harms have emerged, sometimes extremely rapidly, including loneliness and mental health issues. We have been working night and day to address these threats.

Working together with civil society partners, we have launched a dedicated place online that hares NHS advice and other guidance with the admins of Facebook Groups that are discussing COVID-19. People are rallying round to support their neighbours and communities through hundreds of Facebook Groups which have sprung up, staying connected to support each other during these difficult times. We have seen that nearly 2 million people in the UK have now joined more than 2,000 local COVID-19 community support groups on Facebook, and our teams are keeping in touch with thousands of group admins to ensure they have the resources they need to provide their communities with accurate and helpful information.

To further help support the millions of volunteers using Facebook and those in need of support we've launched Community Help - a new hub where people can request or offer help to neighbours such as volunteering to deliver groceries or donating to a local food bank or fundraiser. It's hoped the tool will help connect those self-isolating with much needed supplies as well as giving vulnerable and elderly people a place to ask for help.

We are also working with a series of partners in the Connection Coalition, including the Jo Cox Foundation, Mind and Age UK, aimed at tackling challenges caused by social isolation.We have donated 2,050 Portal devices to the NHS, beginning with a pilot programme in Surrey and expanding across the UK. The devices allow residents in hospitals and care homes to communicate with loved ones that they cannot see during the pandemic.

We are also offering support to Women's Aid for fundraising, and are providing training on how to report concerning content and build a safe community through a Facebook group. We have also worked to support the Home Office and NGOs in the fight against increased rates of domestic violence during the lockdown.

3.4 Each of our services is different, used in different ways for different purposes. Since this crisis began we have been sharing learnings within the business but also investing in tailored approaches for each of Facebook, WhatsApp and Instagram.

WhatsApp is a fundamentally different product from Facebook and Instagram, focused on private sharing between close friends and family - 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. Over the past two years, users have seen a steady drumbeat of changes to our service, all of which have been designed to constrain the virality of content on our service.

In 2018 we began labelling messages that had been forwarded with a single arrow icon. Later that year, we made our first change to our "forward limits", by dropping the number of people that a user can forward a single message to from 20 to just five. This one change for the forward limit alone resulted in a 25 percent decrease in messages that were being forwarded on the service – which roughly translated as one billion fewer messages being forwarded every day.

Last summer we introduced the double arrow label for "highly forwarded" messages—those that have been forwarded more than five times before they reach the user. We recently introduced a new forward limit for highly-forwarded messages on WhatsApp, meaning these messages can only be forwarded to one other contact at a time. This was in part a response to concerns we had seen about misinformation about 5G masts being spread quickly between people on WhatsApp, and the change has resulted in a 70% drop in this high volume forwarding activity.

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. 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.

Specific changes have also been made in relation to COVID-19. We have launched a WhatsApp information hub with tips on how healthcare workers, teachers and local businesses can stay connected using WhatsApp. In partnership with the NHS and Public Health England, we launched our NHS Coronavirus WhatsApp bot, which provides users with interactive answers to questions about the latest figures for the virus, prevention tips and government guidance, and links to mythbusting resources.

Instagram also features tailored approaches to mitigate against harms. When content is rated as false or partly false by an independent third party fact checker we add prominent labels so that it is clear to our community that it has been found to be false or partly false. These labels cover the content itself, whether it is shared in Feed, Stories or Direct Message, and people have to click through to uncover the content. These labels were made more prominent in
October 2019, to give our community more information to help them decide what to read, trust and share. The person who posted the content will also get a notification.

4. The Government's Online Harms proposals

Facebook has for some time welcomed a more active role for the Government in addressing online harms. These are complex issues that we and other companies cannot tackle alone. As a result, we welcomed the Online Harms White Paper and have consistently worked with the UK Government to ensure that the final proposals that flow from any legislation are effective. We support the aim to ensure any new rules for the Internet preserve what's best about it and the incredible benefits it has brought to the daily lives of billions of people, whilst protecting society from broader harms.

We welcome the clear statements in the White Paper that the Government will ensure that the proposals from this document will support the principles of free speech, a vibrant technology sector and will not dampen innovation but in fact place technology at the heart of many of the solutions to document outlines.

We also welcome the statements that both the Government and any future regulator will ensure that such very broad powers will be applied in a proportionate and risk-based way. Overall we are in favour of the "system-based" approach proposed by the White Paper and agree that efforts to regulate content should focus on holding companies accountable for their overall systems to identify and remove harmful content and be backed up by strong transparency obligations.

[COR0178]

We have long believed that empowering people to be digitally savvy is key. This is why we invest in a whole range of tools to give people control over their experience on Facebook - everything from what kind of ads you see to managing your screen time - and we partner with a number of organisations to deliver digital literacy training, safety skills training and resources to young people, parents and teachers.The promise of a coordinated and strategic media literacy strategy from the Government is something we're excited to see.

May 2020