Facebook response to the European Commission Communication on Covid-19 Disinformation

Report for November 2020

1. Introduction

This report builds off our previous reports to the European Commission, in response to the Joint Communication for tackling COVID-19 disinformation, and provides an overview of the policies, products, and processes we have deployed to combat COVID-19 misinformation and disinformation across Facebook and Instagram in November 2020.

2. User Engagement with Authoritative Resources and Tools to Raise Awareness

We continue to find new ways to connect people with accurate, reliable and authoritative information. This is a core component of our strategy to combat misinformation because we want to be able to provide our users with the means to decide what to read, trust and share.

We believe that informing people with accurate and authoritative information, as well as more context, is an approach that can be more impactful than the alternative of just removing content. If we simply removed all posts flagged by fact-checkers as false, for example, the content would still be available elsewhere on the internet, other social media platforms, or even discussed around the dinner table. By leaving this content up and surfacing research from fact-checkers or pointing people to reliable information, we're providing people with important information and context.

As noted by an international group of human rights experts (in relation to COVID-19): "it is essential that governments and internet companies address disinformation in the first instance by themselves providing reliable information. Resorting to other measures, such as content take-downs and censorship, may result in limiting access to important information for public health and should only be undertaken where they meet the standards of necessity and proportionality."

During the coronavirus public health crisis, we have been supporting the global public health community's work to keep people safe and informed by connecting them to accurate, reliable, accessible and relevant sources of information about COVID-19. Our COVID-19 Information Center on Facebook provides people with the latest information from health authorities, news, resources, facts, and tips to stay healthy and safe. It is available globally, including all 27 EU member states. More than 120 million people globally, including over 14 million people in the EU, visited the COVID-19 Information Center during the month of November.

Supporting Local Journalism COVID-19 Information Efforts

The news industry is working under extraordinary conditions to keep people informed during the COVID-19 pandemic. Local journalism is being hit especially hard in the current economic crisis, even as people turn to them for critical information to keep their friends, families and communities safe. Since the start of the pandemic, we have launched several Facebook Journalism Project (FJP) programs to support the news industry during COVID-19, including the COVID-19 Community Network grant program, the COVID-19 News Relief Fund, initiatives via our

Accelerator Program, and the Instagram Local News Fellowship. For Europe specifically, we launched the European Journalism COVID-19 Support Fund, in partnership with the European Journalism Centre (EJC), to help journalists across Europe cover important stories when we all need them most. Here are some examples of how these programs were used to support local journalism:

- **Journalism as a Public Service**: French Regional newspaper company - Groupe Centre France - created a Facebook Messenger chatbot Bonjour Marianne to break down complicated themes (like municipal elections and the local response to COVID-19) in a conversational manner. Group Centre France used the Marianne chatbot to keep audiences engaged, informed and connected during the COVID-19 crisis. Marianne became a hub for support, information and connection when COVID-19 hit, and offered tips for how to make day-to-day activities safer for everyone, spot fake news online and steer clear of information overload. Marianne also collected feedback on user behaviour, quality of life, mental and emotional well-being to gain insight from readers on how local news outlets can support the public during a crisis. This change in direction for the chatbot attracted a whole new audience, exponentially growing Groupe Centre France's readership, and showed how talking and listening to readers allows publishers to better serve them.

- **Engaging Audiences on COVID-19 Coverage**: La Calle TV, a digital-first news organization that delivers culturally relevant stories to a growing Latin audience in the US, became one of the most important news outlets in New York City for its wall-to-wall COVID-19 coverage, as their Facebook Page saw 20x more interactions during the peak of the pandemic (March–April 2020) compared to the two months prior. Through frequent video updates on Facebook, La Calle TV gained a following of Latin US citizens looking for reliable news on COVID-19.

- **Elevating Hyperlocal Storytelling About COVID-19**: The Instagram Local News Fellowship paired 22 budding journalists with 21 local newsrooms across the US in the summer of 2020. The fellows, all recent college graduates or college seniors, were tasked with using their newsrooms' platforms to connect, serve and engage their local communities with information pertaining to the spread of COVID-19, as well as amplifying safety guidelines as they emerged from local and federal officials. Through Instagram and a series of IGTV videos, the fellows were able to elevate their storytelling to a national level, shining a much needed spotlight on hyperlocal reporting.

3. Actions on Misinformation

Our goal is to create a place for expression and give people a voice. Building community and bringing the world closer together depends on people's ability to share diverse views, experiences, ideas and information. Our commitment to expression is paramount, but we recognise that the internet creates new and increased opportunities for abuse. When considering whether to provide more context, allow, reduce distribution, or remove misinformation, we do it in service of one or more of our Community Standards Values: voice, authenticity, safety, privacy, and dignity.

We define misinformation as content that is false or misleading. We enforce on misinformation by looking at content or behaviors that violate our Community Standards or content that may be reviewed by our third-party fact-checking partners. We define disinformation as coordinated efforts to manipulate public debate for a strategic goal.

Applying Community Standards to COVID-19 Content

As people around the world confront this unprecedented public health emergency, we want to make sure that our Community Standards protect people from harmful content and new types of abuse related to COVID-19. We're working to remove content that has the potential to contribute to offline harm, including through our policies prohibiting the coordination of harm, the sale of medical masks and related goods, hate speech, bullying and harassment, and misinformation that contributes to the risk of imminent violence or physical harm. Oftentimes, misinformation can cut across different types of abuse areas; for example, a racial slur could be coupled with a false claim about a group of people and we'd remove it for violating our hate speech policy. So in addition to our misinformation policies, we have a number of other ways we might combat COVID-19 misinformation such as:

- Under our **Regulated Goods** policy, we've taken steps to protect against exploitation of this crisis for financial gain by banning content that attempts to sell or trade medical masks, hand sanitizer, surface disinfecting wipes and COVID-19 test kits. We also prohibit influencers from promoting these sales through branded content. During the month of November, we removed over 240 thousand pieces of content on Facebook and Instagram globally, including over 10 thousand pieces of content in the EU member states, related to COVID-19 and which violated our medical supply sales standards.

- Under our **Hate Speech** policy, we are removing content that states that people who share a protected characteristic such as race or religion have the virus, created the virus or are spreading the virus. This does not apply to claims about people based on national origin because we want to allow discussion focused on national-level responses and effects (e.g., "X number of Italians have COVID-19"). We also remove content that mocks people who share a protected characteristic such as race or religion for having COVID-19. As reported in our Community Standards Enforcement Report (CSER):

    - Facebook: Content actioned and the proactive rate for hate speech remained similar across Q2 2020 and Q3 2020. We took action on 22.1 million pieces of content globally from July to September 2020. Prevalence of hate speech content, which we published for the first time in the latest report, was between 0.10% and 0.11% of views in Q3.

    - Instagram: Content actioned increased from 3.2 million pieces of content in Q2 2020 to 6.5 million in Q3 2020, partly due to expanding automation to the Arabic and Indonesian languages toward the end of Q2, which continued to drive enforcement in Q3. We followed up in Q3 by improving our proactive detection technology for the English, Arabic and Spanish languages, and expanded automation for violating media and comments, which helped drive the increase in our proactive rate from 84.9% to 94.8%.

- Under our **Bullying and Harassment** policy, we remove content that targets people maliciously, including content that claims that a private individual has COVID-19, unless that person has self-declared or information about their health status is publicly available. As reported in our Community Standards Enforcement Report (CSER):

    - Facebook: Content actioned increased from 2.4 million pieces of content in Q2 2020 to 3.5 million in Q3 2020. This was partly due to an issue with our proactive detection technology that caused us to mistakenly remove non-violating comments,

which we later restored. We also increased our automation abilities beginning in Q2. Our proactive rate increased from 13.3% to 26.4% for these same reasons.

  - Instagram: Content actioned increased from 2.3 million pieces of content in Q2 2020 to 2.6 million in Q3 2020. This was partly due to an issue with our proactive detection technology that caused us to mistakenly remove non-violating comments, which we later restored. We also made improvements to our proactive detection technology in Q2, which continued to drive enforcement in Q3. Our proactive rate increased from 37.7% to 54.7% for these same reasons.

- Under our **Misinformation and Harm** policy, we remove misinformation that contributes to the risk of imminent violence or physical harm. We have applied this policy to harmful misinformation about COVID-19 since January. During the month of November, we removed over 360 thousand pieces of content on Facebook and Instagram globally, including over 35 thousand pieces of content in the EU, for containing misinformation that may lead to imminent physical harm, such as content relating to fake preventative measures or exaggerated cures.

As the situation evolves, we are continuing to look at content on the platform, assess speech trends and engage with experts, and will provide additional policy guidance to our Community Standards when appropriate to keep the members of our community safe during this crisis. These policies, as well as the additional policies listed in our Community Standards apply to content on both Facebook and Instagram, including surfaces such as Groups and Pages.

Using AI to Help Detect Misinformation and Deepfakes

Artificial Intelligence is a critical tool to help protect people from harmful content. It helps us scale the work of human experts, and proactively take action, before a problematic post or comment has a chance to harm people.

Facebook has implemented a range of policies and products to deal with misinformation on our platform. These include adding warnings and more context to content rated by third-party fact-checkers, reducing their distribution, and removing misinformation that may contribute to imminent harm. But to scale these efforts, we need to quickly spot new posts that may contain false claims and send them to independent fact-checkers — and then work to automatically catch the same content, so fact-checkers can focus their time and expertise fact-checking new content. Our AI tools both flag likely problems for review and automatically find new instances of previously identified misinformation.

As with hate speech, this poses difficult technical challenges. Two pieces of misinformation might contain the same claim but express it very differently, whether by rephrasing it, using a different image, or switching the format from graphic to text. And since current events change rapidly, especially in the run-up to an election, a new piece of misinformation might focus on something that wasn't even in the headlines the day before.

To better apply warning labels at scale, we are developing new AI technologies to match near-duplications of known misinformation at scale. When fact-checkers have identified a piece of misinformation, our AI would help spot copies of cropped or altered content and content that convey the same meaning but look different.

We've also taken steps to deal with deepfakes, which use AI to show people doing and saying things they didn't actually do or say and can be difficult for even a trained reviewer to spot. Our AI Red team ran experiments to help anticipate potential problems and we have now deployed a state-of-the-art deepfake detection model, which uses multiple generative adversarial networks (GANs) to train our detection system. This work was the result of the Deepfake Detection Challenge (DFDC), which is an open, collaborative initiative organized by Facebook and other industry leaders and academic experts.

We're making progress, but we know our systems are far from perfect and there's much more work to do. For more information on our AI misinformation and deepfakes work, see here.

Our Third-Party Fact-Checking Program

For misinformation that does not lead to real world harm, but undermines the authenticity and integrity of our platform, we continue to work with our growing network of independent third party fact-checking partners. We partner with nearly 80 fact-checking organizations around the world, covering over 60 languages. In the EU and greater Europe, we expanded our program into Romania with AFP. This brings the total number of fact-checkers for the region to 36, covering 26 languages.

Based on the work of our fact-checking partners, we displayed misinformation warning screens associated with COVID-19 related fact-checks on over 22 million pieces of content globally, including over 3.4 million pieces of content in EU member states, in November.

Community Standards Enforcement Report

In November, we published our Community Standards Enforcement Report (CSER) for the third quarter of 2020. This report provides metrics on how we enforced our policies globally from July through September and includes metrics across 12 policies on Facebook and 10 policies on Instagram, including Adult Nudity and Sexual Activity, Bullying and Harassment, Child Nudity and Sexual Exploitation of Children, Dangerous Organizations: Terrorism and Organized Hate, Fake Accounts, Hate Speech, Regulated Goods: Drugs and Firearms, Spam, Suicide and Self-Injury, and Violent and Graphic Content.

For the first time in our CSER report, we included hate speech prevalence on Facebook globally. In Q3 2020, hate speech prevalence was 0.10% – 0.11% or 10 to 11 views of hate speech for every 10,000 views of content. Due to our investments in AI, we have been able to remove more hate speech and find more of it proactively before users report it to us. Our enforcement metrics Q3, including how much hate speech content we found proactively and how much content we took action on, indicate that we're making progress catching harmful content. Prevalence, on the other hand, estimates the percentage of times people see violating content on our platform. Read more about our work on hate speech and how AI is getting better at detecting hate speech.

While the COVID-19 pandemic continues to disrupt our content review workforce, we are seeing some enforcement metrics return to pre-pandemic levels. Our proactive detection rates for violating content are up from Q2 across most policies, due to improvements in AI and expanding our detection technologies to more languages. Even with a reduced review capacity, we still prioritize the most sensitive content for people to review, which includes areas like suicide and self-injury and child nudity. Here are some key metrics in the CSER report:

On Facebook in Q3, we took action on:

- 22.1 million pieces of hate speech content, about 95% of which were proactively identified
- 19.2 million pieces of violent and graphic content (up from 15 million in Q2)
- 12.4 million pieces of child nudity and sexual exploitation content (up from 9.5 million in Q2)
- 3.5 million pieces of bullying and harassment content (up from 2.4 million in Q2)
- 1.3 billion accounts, 99.3% of which we found and flagged before users reported them

On Instagram in Q3, we took action on:

- 6.5 million pieces of hate speech content (up from 3.2 million in Q2), about 95% of which was proactively identified (up from about 85% in Q2)
- 4.1 million pieces of violent and graphic content (up from 3.1 million in Q2)
- 1.0 million pieces of child nudity and sexual exploitation content (up from 481,000 in Q2)
- 2.6 million pieces of bullying and harassment content (up from 2.3 million in Q2)
- 1.3 million pieces of suicide and self-injury (up from 277,400 in Q2)

The increase in our proactive detection rate for hate speech on Instagram was driven in part by improving our proactive detection technology for English, Arabic and Spanish languages, and expanding automation technology. We expect fluctuations in these numbers as we continue to adapt to COVID-related workforce challenges.

We also updated our Community Standards website to include additional policies that require more context and can't always be applied at scale. Several of these policies have been announced before. For example, our policy that prohibits posting misinformation and unverifiable rumors that contribute to the risk of imminent violence or physical harm, and our policy to add a warning label to sensitive content such as imagery posted by a news agency that depicts child nudity in the context of famine, genocide, war crimes or crimes against humanity. While these policies are not new, we are sharing more details to be even more transparent about our enforcement practices. Moving forward, just as we do with our scaled policies, we will continue to publicly update the Community Standards monthly as new policies are developed that require additional context.

4. Coordinated Inauthentic Behaviour (CIB) and Influence Operations

We know that people looking to mislead others - whether through phishing, scams, or influence operations - try to leverage crises in order to advance their goals, and the COVID-19 pandemic is no different. As the situation evolves, we are actively working to find and stop coordinated campaigns that seek to manipulate public debate across our platforms.

Our approach to Coordinated Inauthentic Behavior (CIB), and Influence Operations (IO) more broadly, is grounded on behavior- and actor-based enforcement. This means that we are looking for specific violating behaviours exhibited by violating actors, rather than violating content (which is predicated on other specific violations of our Community Standards, such as misinformation and hate speech). Therefore, when CIB networks are taken down, it is based on their behavior, not the content they posted. For a comprehensive overview of our approach, see here.

To date, we have not found evidence of influence operations created to focus specifically on COVID-19. What we have seen is that people behind campaigns opportunistically use coronavirus-related posts among many other topics to build an audience and drive people to their Pages or off-platform sites.

Detailed reports on the CIB networks taken down and examples of content they posted can be found here.

Taking Action Against Hackers

Facebook's threat intelligence analysts and security experts work to find and stop a wide range of threats including malware campaigns, influence operations and hacking of our platforms or individual Facebook accounts by nation state adversaries, hackers and others. As part of these efforts, our teams routinely disrupt adversary operations by disabling them, notifying users if they should take steps to protect their accounts, sharing our findings publicly and continuing to improve the security of our products.

For the first time, in November, we shared our latest research and enforcement actions against attempts to compromise people's accounts and gain access to their information, commonly referred to as cyber espionage. Although this first report focused on operations in Asia, it provides insights on how these tactics are evolving and how Facebook is responding. These operations may leverage certain crises or critical events, such as the COVID-19 pandemic, to advance their goals, so removing them from our platforms is another critical step towards keeping our platforms safe and secure for our community around the world.

We took action against two separate groups of hackers - APT32 in Vietnam and a group based in Bangladesh - removing their ability to use their infrastructure to abuse our platform, distribute malware and hack people's accounts across the internet. These two unconnected groups targeted people on our platform and elsewhere on the internet using very different tactics. The operation from Vietnam focused primarily on spreading malware to its targets, whereas the operation from Bangladesh focused on compromising accounts across platforms and coordinating reporting to get targeted accounts and Pages removed from Facebook.

Bangladesh: The Bangladesh-based group targeted local activists, journalists and religious minorities, including those living abroad, to compromise their accounts and have some of them disabled by Facebook for violating our Community Standards. Our investigation linked this activity to two non-profit organizations in Bangladesh: Don's Team (also known as Defense of Nation) and the Crime Research and Analysis Foundation (CRAF). They appeared to be operating across a number of internet services. Don's Team and CRAF collaborated to report people on Facebook for fictitious violations of our Community Standards, including alleged impersonation, intellectual property infringements, nudity and terrorism. They also hacked people's accounts and Pages, and used some of these compromised accounts for their own operational purposes, including to amplify their content. On at least one occasion, after a Page admin's account was compromised, they removed the remaining admins to take over and disable the Page. Our investigation suggests that these targeted hacking attempts were likely carried out through a number of off-platform tactics including email and device compromise and abuse of our account recovery process. We removed the accounts and Pages behind this operation, and shared information about this group with our industry partners so they too can detect and stop this activity.

Vietnam: APT32, an advanced persistent threat actor based in Vietnam, targeted Vietnamese human rights activists locally and abroad, various foreign governments including those in Laos and Cambodia, NGOs, news agencies and a number of businesses across information technology, hospitality, agriculture and commodities, hospitals, retail, the auto industry, and mobile services with malware. Our investigation linked this activity to CyberOne Group, an IT company in Vietnam

(also known as CyberOne Security, CyberOne Technologies, Hành Tinh Company Limited, Planet and Diacauso). As our industry partners have previously reported, APT32 has deployed a wide range of adversarial tactics across the internet. We have been tracking and taking action against this group for several years. Our most recent investigation analyzed a number of notable tactics, techniques and procedures (TTPs) including:

- Social engineering: APT32 created fictitious personas across the internet posing as activists and business entities, or used romantic lures when contacting people they targeted. These efforts often involved creating backstops for these fake personas and fake organizations on other internet services so they appear more legitimate and can withstand scrutiny, including by security researchers. Some of their Pages were designed to lure particular followers for later phishing and malware targeting.

- Malicious Play Store apps: In addition to using Pages, APT32 lured targets to download Android applications through Google Play Store that had a wide range of permissions to allow broad surveillance of peoples' devices.

- Malware propagation: APT32 compromised websites and created their own to include obfuscated malicious javascript as part of their watering hole attack to track targets' browser information. A watering hole attack is when hackers infect websites frequently visited by intended targets to compromise their devices.

The latest APT32 activity we investigated and disrupted has the hallmarks of a well-resourced and persistent operation focusing on many targets at once, while obfuscating their origin. We shared our findings including YARA rules and malware signatures with our industry peers so they too can detect and stop this activity. We blocked associated domains from being posted on our platform, removed the group's accounts and notified people who we believe were targeted by APT32.

People behind these cyber espionage operations are persistent adversaries, and we expect them to evolve their tactics. However, our detection systems and threat investigators, as well as other teams in the security community, keep improving to make it harder for them to remain undetected. We will continue to share our findings whenever possible so people are aware of the threats we are seeing and can take steps to strengthen the security of their accounts. You can find more information on our latest research and actions here.

5. Advertising

As the COVID-19 situation develops, we have implemented a variety of measures to prevent ads from being used to spread misinformation; to prevent ads from promoting content that could contribute to physical harm; to prohibit exploitative or deceptive ads; and provide transparency on ads about health issues. We have applied our Advertising Policies to new types of abuse that we're seeing on the platform. We have made adjustments to our enforcement protocols to prevent people from exploiting the COVID-19 pandemic and continue adapting or removing temporary bans on specific products as the situation stabilizes. For a full list of our Advertising Policies about COVID-19, see here.