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

Report for August 2020

1. Introduction

This report builds off of our previous reporting to the European Commission on how we combat misinformation, in response to the Joint Communication for tackling COVID-19 disinformation. This report focuses on the policies, products, and processes we have deployed to tackle COVID-19 misinformation across Facebook and Instagram in August 2020.

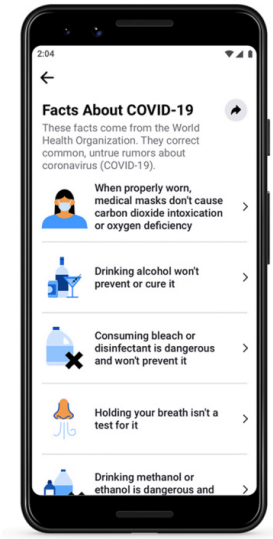
Our previous report, covering the period of January to July 2020, can be found here.

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

Empowering people to decide for themselves what to read, trust, and share is an important part of our strategy to combat misinformation. During the coronavirus public health crisis, we are supporting the global public health community's work to keep people safe and informed. Partnering with global and local health authorities, we are working to connect people to accurate, authoritative, accessible and relevant sources of information about COVID-19. Our COVID-19 Information Center on Facebook continues connecting people to the latest information from health authorities, news, resources, facts, and tips to stay healthy and safe. It is available in 189 countries, including all 27 EU member states. More than 13 million people in the EU visited the COVID-19 Information Center during the month of July, and more than 14 million people during the month of August.

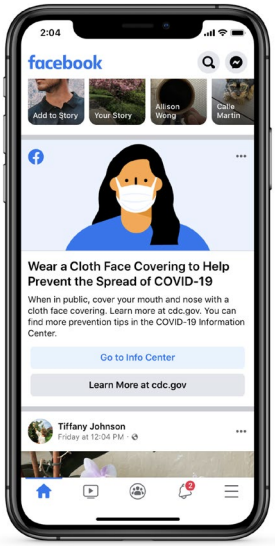
Facts About COVID-19

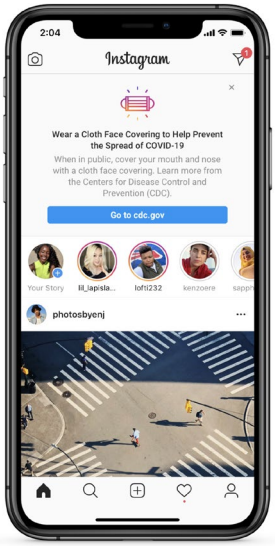
To further limit the spread of misinformation, we launched a dedicated section in the COVID-19 Information Center called Facts about COVID-19. It debunks common myths that have been identified by the World Health Organization such as drinking bleach will prevent the coronavirus or that holding your breath for 10 seconds without coughing means you don't have coronavirus. This is the latest step in our ongoing work to fight misinformation about the pandemic.



Global Reminders to Wear Face Coverings

With the rise in COVID-19 cases in many parts of the world, we started putting an alert at the top of Facebook and Instagram to remind people to wear face coverings, as recommended by health authorities. These alerts started running at the top of Facebook and Instagram since early July, and in the EU has been launched in Austria, Belgium, Bulgaria, Croatia, Cyprus, Czech Republic, Estonia, Finland, France, Germany, Greece, Hungary, Ireland, Italy, Latvia, Lithuania, Luxembourg, Malta, Poland, Portugal, Romania, Slovakia, Spain and Slovenia.





Recommendations Guidelines

Across our apps, we make recommendations to help people discover new communities and content we think they would likely be interested in. We suggest Pages, Groups, Events and more based on content people have expressed interest in and actions they take on our apps. We personalize these recommendations to make sure they are relevant and valuable to each individual.

Recommendations can help people discover things they love, but since recommended content does not come from accounts people choose to follow, it is important that we have certain

standards for what we recommend. This helps ensure we don't recommend potentially sensitive content to those who don't explicitly indicate that they wish to see it. To be clear, this content is still allowed on our platforms – we just won't show it in places where we recommend content.

To determine what content is eligible to appear in recommendations, we have Recommendation Guidelines. We have now made these guidelines public in the Help Center to help people better understand the kinds of content we recommend, and provide context on why some types of content aren't included in recommendations, and therefore may not be distributed as widely. In developing these guidelines, we consulted 50 leading experts specializing in recommender systems, expression, safety and digital rights. These consultations help ensure we provide a safe and positive experience when we recommend things on our apps. The Recommendations Guidelines are available in the Facebook Help Center and Instagram Help Center.

Resources for Governments Responding to COVID-19

We know that keeping people safe and informed is an important function of government organisations during the COVID-19 public health crisis. Providing people with reliable and timely information to help keep them safe and to strengthen community is often key to overcoming a crisis. For governments in particular, crises such as the coronavirus (COVID-19) pandemic has emphasised the need for swift, clear and direct communication with citizens.

While social media enables government organisations to reach a large audience quickly with real-time information during a crisis, we realise that it's not always easy to manage. From knowing the range of tools available to best practices around content and community management, there is much to consider. To help governments communicate effectively in times of crisis, we have compiled information and resources in a COVID-19 Hub for Governments that helps them get the best out of Facebook apps and services to reach people with relevant and timely information.

It includes a comprehensive Crisis Communication Guide and video featuring insights on tools, best practices and creative ways for government organisations to harness Facebook apps and services as they keep people informed during a crisis.

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.

Our approach to misinformation is grounded in content-based enforcement. This means that we are looking at content that violates our Community Standards or content that may be reviewed by our third-party fact-checkers.

Applying Community Standards to COVID-19 Content

We continue reviewing and ensuring that our Community Standards protect people from harmful content and new types of abuse related to COVID-19. We remove content that has the potential to contribute to real-world harm, according to our policies for misinformation that contributes to the risk of imminent violence or physical harm; prohibiting the coordination of harm; the sale of regulated goods; hate speech; and bullying and harassment.

- 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. In July, we removed more than 31 thousand pieces of content on Facebook and Instagram in the EU for containing misinformation that may lead to imminent physical harm, such as content relating to fake preventative measures or exaggerated cures. In August, we removed more than 36 thousand pieces of such content on Facebook and Instagram in the EU.

- Under our **Coordinating Harm** policy, we remove content that advocates for the spread of COVID-19 as well as content that encourages or coordinates the physical destruction of infrastructure, such as 5G masts. This also includes removing content coordinating in-person events or gatherings when participation involves or encourages people with COVID-19 to join.

- 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 July, we removed at least 13 thousand pieces of content from Facebook and Instagram in EU member states related to COVID-19 and which violated our medical supply sales standards. During the month of August, we removed at least 15 thousand pieces of such content from Facebook and Instagram in the EU.

- Under our **Hate Speech** policy, we remove 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) that was released in August, the amount of hate speech content we took action on increased from 9.6 million in Q1 2020 to 22.5 million in Q2 2020. These figures include COVID-19 related content. We made improvements to our proactive detection technology and expanded automation to more languages which helped us detect and remove more content.

- 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) that was released in August, the amount of content we removed for violating our bullying and harassment policy increased from 2.3 million in Q1 2020 to 2.4 million in Q2 2020.

These figures include COVID-19 related content. After enforcement was impacted by temporary workforce changes due to COVID-19, we regained some review capacity in Q2. We also increased our automation abilities and made improvements to our proactive detection technology for the English language.

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.

New Ratings for Fact-Checking Partners

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 now partner with over 70 fact-checking organizations around the world, covering over 60 languages. This includes the 34 fact-checking partners that we have in the EU and greater Europe, covering 24 languages. Based on the work of our fact-checking partners, we displayed misinformation warning screens associated with COVID-19 related fact-checks on over 4.1 million pieces of content in EU member states in July. We displayed misinformation warning screens associated with COVID-19 related fact-checks on over 4.6 million pieces of such content in EU member states in August.

We continue to evolve our fact-checking program to reflect the content our partners see on our platforms. Our approach is to take strong action on the worst of the worst misinformation, while also accounting for content that might not be outright false, but could benefit from additional context. To that effect, we recently announced two new ratings to provide our fact-checking partners with more latitude to better reflect their research, and to help ensure that people who come across these posts have more precise information to judge what to read, trust, and share.

- **“Altered” Content:** The first rating, called “Altered,” is designed specifically for videos and images that have been manipulated in ways that could mislead people. For example, an edited video that shows someone shaking someone’s hand when they didn’t. Or an image where someone used Photoshop to depict a person at a location that they weren’t actually at. (We also remove certain manipulated media under our Community Standards).

- **“Missing Context”:** The second rating, called “Missing Context,” is designed for content that may mislead without additional context. Over the past few months, our fact-checking partners have increasingly flagged examples of this type of content to us, which wasn’t accurately represented by our existing ratings. For example, cropping a video clip to take out certain words; changing it from: “I would support this candidate if…” to saying “I support this candidate.” Or, claiming that funding for a government program has been “zeroed out” when its funding has been dramatically reduced but not eliminated.

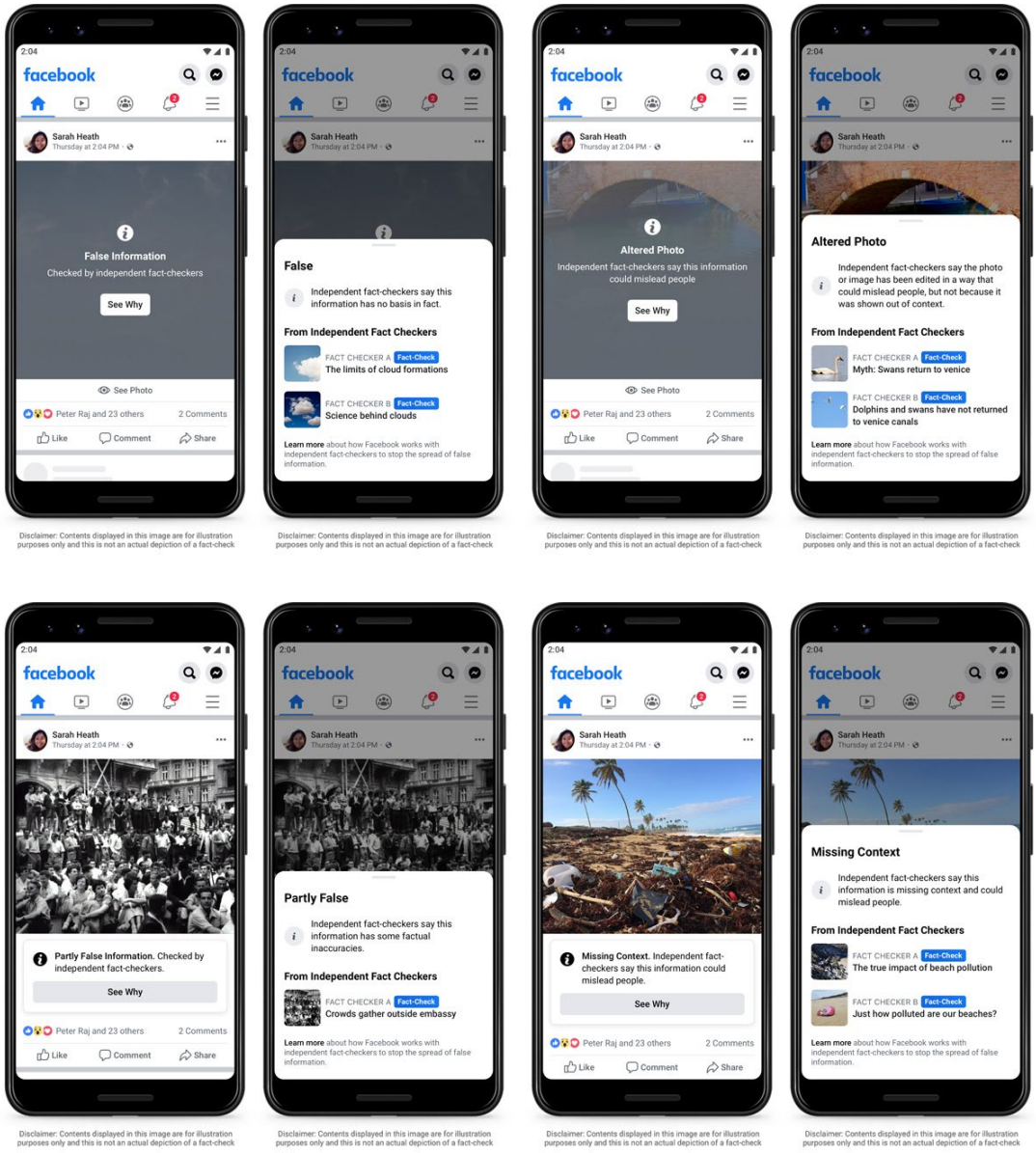
For both ratings, we'll be introducing a lighter-weight warning label to more precisely reflect fact-checkers' assessments.

Content rated either **“False”** or **“Altered”** makes up the worst of the worst kind of misinformation. As such, these ratings will result in our most aggressive actions: we will dramatically reduce the distribution of these posts, and apply our strongest warning labels. Content rated **“Partly False”** includes some factual inaccuracies. As a result, we reduce the

distribution of this content, but to a lesser degree than "False" or "Altered." For content that's "Missing Context," we'll focus on surfacing more information from our fact-checking partners.

We also recently clarified some confusion about our approach to opinion. This content is generally not eligible for fact-checking because we don't want to interfere with individual expression. But there is an important exception. If content is presented as opinion but is based on underlying false information - even if it's an op-ed or editorial - it's still eligible to be fact-checked. We are doing this because presenting something as opinion isn't meant to give a free pass to content that spreads false information.

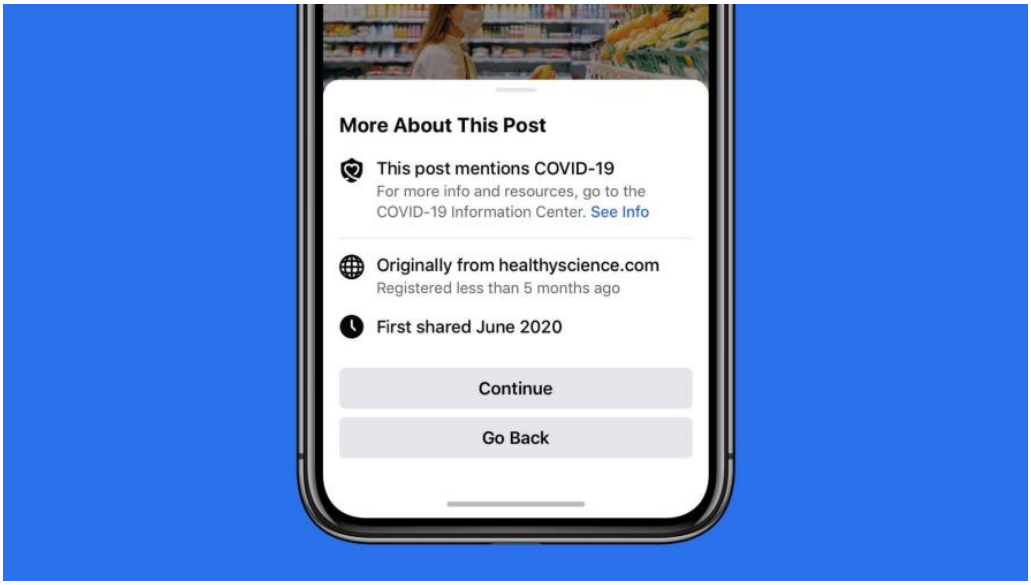
During the COVID-19 pandemic and heading into the elections in the U.S. and around the world, we realize how important it is for people to understand what they're seeing when they're using our services and then judge its worth. We'll begin to roll these new ratings out globally throughout the coming weeks.



Giving People More Context About COVID-19 Links

We want to make sure people have the context they need to make informed decisions about what to share on Facebook, especially when it comes to COVID-19 content. Building on the notification screen that is displayed when people share articles older than 90 days, which we launched in June, we have started to roll out a global notification screen to give people more context about COVID-19 related links when they are about to share them.

The notification will help people understand the recency and source of the content before they share it. It will also direct people to our COVID-19 Information Center to ensure people have access to credible information about COVID-19 from global health authorities. Along those lines, we want to ensure we don't slow the spread of information from credible health authorities, so content posted by government health authorities and recognized global health organizations, like the World Health Organization, will not have this notification.



Media Literacy Campaign to Help Spot False News

We want to give people the tools to make informed decisions about the information they see online and where it comes from. To support this effort, we ran a media literacy campaign in June and August in 56 countries (including all 27 EU member states) and 27 languages to raise awareness and educate people about how to spot potential false news. In consultation with some of our fact-checking partners, we developed 'Three questions to help stamp out false news'. More than 1.3 million people in the EU clicked on the campaign and visited the website to learn more about how to identify false news.

Introducing New Authenticity Measures on Instagram

We want the content that people see on Instagram to be authentic and to come from real people, not bots or others trying to mislead. Starting August, we began asking people to confirm who's behind an account when we see a pattern of potential inauthentic behavior. By prompting the people behind accounts to confirm their information, we will be able to better understand

when accounts are attempting to mislead their followers, hold them accountable, and keep our community safe.

We will look at a range of signals to determine if an account holder needs to confirm their information. We want to be clear that this change will apply only to a small number of our community. Most people will not be affected. This includes accounts potentially engaged in coordinated inauthentic behavior, or when we see the majority of someone's followers are in a different country to their location, or if we find signs of automation, such as bot accounts for example. If we see signs of potential inauthentic activity, we will require the account holder to confirm who they are, and once an account holder verifies their information, their account will function as usual unless we have reason to investigate further. IDs will be stored securely and deleted within 30 days once our review is completed, and won't be shared on the person's profile as pseudonymity is still an important part of Instagram.

If an account chooses not to confirm their information, their content may receive reduced distribution, or the account may be disabled. Visit the Help Center for more information on the identity verification process, and the types of IDs we accept.

Addressing Movements and Organizations Tied to Violence

We are now taking action against Facebook Pages, Groups and Instagram accounts tied to offline anarchist groups that support violent acts amidst protests, US-based militia organizations and QAnon. We already remove content calling for or advocating violence and we ban organizations and individuals that proclaim a violent mission. However, we have seen growing movements that, while not directly organizing violence, have celebrated violent acts, shown that they have weapons and suggest they will use them, or have individual followers with patterns of violent behavior. So we have expanded our Dangerous Individuals and Organizations policy to address organizations and movements that have demonstrated significant risks to public safety but do not meet the rigorous criteria to be designated as a dangerous organization and banned from having any presence on our platform. While we will allow people to post content that supports these movements and groups, so long as they do not otherwise violate our content policies, we will restrict their ability to organize on our platform.

Under this policy expansion, we will impose restrictions to limit the spread of content from Facebook Pages, Groups and Instagram accounts. We will also remove Pages, Groups and Instagram accounts where we identify discussions of potential violence, including when they use veiled language and symbols particular to the movement to do so. For more information on this update, see here.

Funding for Research on Misinformation

Misinformation and polarization are fundamental challenges we face, not just as a company with the mission of bringing people together but also as members of societies dealing with layered challenges ranging from election interference to a global pandemic.

At the end of February, Facebook Research launched a request for proposals focusing on these dual challenges. Our goal is to support independent research that will contribute to the understanding of these phenomena and, in the long term, help us improve our policies, interventions, and tooling. We invited proposals that took any of a wide variety of research

approaches to bring new perspectives into ongoing work on issues like health misinformation, affective polarization, digital literacy, and more.

We received over 1,000 proposals from 600 institutions and 77 countries around the world that covered an impressive range of disciplines and methodological approaches. We have selected 25 awardees, who will investigate these issues across 42 countries: Argentina, Australia, Brazil, Canada, Chile, China, Colombia, Denmark, Egypt, Ethiopia, Germany, Ghana, Hungary, India, Indonesia, Israel, Italy, Kenya, Mexico, Myanmar, New Zealand, Nigeria, Pakistan, Philippines, Russia, Rwanda, Spain, South Africa, South Korea, Sudan, Taiwan, Tanzania, Thailand, Tunisia, Turkey, Uganda, Ukraine, United Arab Emirates, United Kingdom, United States, Vietnam, and Zimbabwe. The full list of awardees can be found here.

Independent Audit of Community Standards Enforcement Report Metrics

To track our progress and demonstrate our continued commitment to making Facebook and Instagram safe and inclusive, we regularly release the Community Standards Enforcement Report (CSER). This report shares metrics on how we are doing at preventing and taking action on harmful content that goes against our Community Standards. Facebook works with a wide range of stakeholders, academics, and experts to help us assess what the right metrics to share in the report are.

One of the most robust efforts in this space was our work with the Data Transparency Advisory Group (DTAG) — a group comprised of international experts in measurement, statistics, criminology, and governance — to provide an independent, public assessment of whether the metrics we share in the Community Standards Enforcement Report are meaningful and accurate. We provided the advisory group detailed and confidential information about our enforcement processes and measurement methodologies. DTAG subsequently published an independent assessment, which found our approach and methodology sound and reasonable, and noted that the CSER is an important exercise in transparency.

In August, we issued a Request for Proposal to external auditors to conduct an independent audit of our CSER metrics. This unprecedented audit aims to give external stakeholders, including regulators, additional confidence that the numbers we are reporting around harmful content in the CSER are accurate. We hope to conduct this audit starting in 2021 and have the auditors publish their assessments once completed. No company should grade its own homework and the credibility of our systems should be earned, not assumed. We believe independent audits and assessments are crucial to hold us accountable and help us do better.

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 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.

August 2020 Coordinated Inauthentic Behavior Report

In August, we removed three networks of accounts, Pages and Groups. Two of them — from Russia and the US — targeted people outside of their country, and another from Pakistan focused on both domestic audiences in Pakistan and also in India. We have shared information about our findings with law enforcement, policymakers and industry partners.

Since 2017, we have removed over 100 networks worldwide for engaging in coordinated inauthentic behavior, including ahead of major democratic elections. The first network we took down was linked to the Russian Internet Research Agency (IRA), and so was the 100th we took down in August. In total, our team has found and removed about a dozen deceptive campaigns connected to individuals associated with the IRA. Over the last three years, we have detected these efforts earlier and earlier in their operation, often stopping them before they were able to build their audience. With each takedown, threat actors lose their infrastructure across many platforms, forcing them to adjust their techniques, and further reducing their ability to reconstitute and gain traction.

As part of our work to find, study and remove influence operations from Facebook, we've seen them target multiple technology platforms and seek to use traditional media to amplify their narratives. We've seen a number of campaigns, including the two we removed in August, create Pages purporting to be news entities to appear more credible. The IRA-linked campaign we removed in August was largely unsuccessful on Facebook, but it tricked unwitting freelance journalists into writing stories on its behalf. We're notifying people who we believe have been contacted by this network.

We expect to see more attempts like this from threat actors globally and we'll remain vigilant and work with other technology companies, law enforcement, and independent researchers to find and remove influence operations.

The networks reported below were removed for behaviours that violated our Inauthentic Behaviour Policy. As noted, we have not found evidence of COVID-19 focused influence operations.

- Total number of Facebook accounts removed: 521
- Total number of Instagram accounts removed: 72
- Total number of Pages removed: 147
- Total number of Groups removed: 78

(Note: These numbers may be updated when more data for this reporting period becomes available.)

Networks removed in August 2020:

1. **Russia:** We removed a small network of 13 Facebook accounts and two Pages linked to individuals associated with past activity by the Russian Internet Research Agency (IRA). This activity focused primarily on the US, UK, Algeria and Egypt, in addition to other English-speaking countries and countries in the Middle East and North Africa. We began this investigation based on information about this network's off-platform activity from the FBI. Our internal investigation revealed the full scope of this network on Facebook.

2. **US:** We removed 55 Facebook accounts, 42 Pages and 36 Instagram accounts linked to US-based strategic communications firm CLS Strategies. This network focused primarily on Venezuela and also on Mexico and Bolivia. We found this activity as part of our proactive investigation into suspected coordinated inauthentic behavior in the region.

3. **Pakistan:** We removed 453 Facebook accounts, 103 Pages, 78 Groups and 107 Instagram accounts operated from Pakistan and focused on Pakistan and India. We found this network as part of our internal investigation into suspected coordinated inauthentic behavior in the region.

We are making progress rooting out this abuse, but it's an ongoing effort. We're committed to continually improving to stay ahead. That means building better technology, hiring more people and working closely with law enforcement, security experts and other companies.

A detailed report on the networks taken down and examples of content they posted can be found here. Previous reports can be found 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.

Allowing the Promotion and Sale of Hand Sanitizer and Surface Disinfecting Wipes

In March, we temporarily banned ads and commerce listings for hand sanitizer and surface disinfecting wipes to help protect against scams, inflated prices and hoarding. Since then, we've continued to monitor trends and activity around COVID-19 to better understand how people are using our platform and advertising tools during the pandemic. In August, we scaled back this temporary ban to allow people to promote and trade hand sanitizer and surface disinfecting wipes on our apps.

For a full list of our Advertising Policies about COVID-19, see here.

Our previous report, covering the period of January to July 2020, can be found here.