ISD

Powering solutions
to extremism
and polarisation

The First 100 Days: Coronavirus and Crisis Management on Social Media Platforms

Chloe Colliver & Jennie King

About this paper

This report offers an interim review of responses to the COVID-19 'infodemic' from three major technology companies - Facebook, Google and Twitter - from March to May 2020. The report summarises the approaches taken by respective teams at Twitter, Facebook, WhatsApp, Instagram, Google and YouTube, including specific services and policies introduced in recent months and, where possible, the accompanying rationale from companies themselves. The report reviews research from ISD and additional organisations to assess the enforcement and efficacy of these policies, and sets out six recommendations to improve defence against disinformation in the future.

Acknowledgements

This report was written with support from Henry Tuck and Milo Comerford.

ISD wishes to thank all the organisations whose research contributed to this assessment and who continue to do work to expose, analyse and respond to disinformation.

Contents

Executive Summary

Since January 2020, COVID-19 has become the perfect crucible for online harms. Pandemics are by their nature fast-moving, with constantly evolving information even from credible and expert sources. This is set against a backdrop of heightened fear and anxiety, where valid concerns over resource scarcity, economic fallout and personal safety merge with extremist views on race and social order. New conspiracies and coordinated disinformation efforts have exploded online, preying on the uncertainty of this moment and the ambiguity regarding the source and spread of the disease worldwide.

The disinformation crisis surrounding COVID-19 is not an abstract problem. Online content can catalyse real-world harm, and research is already documenting the risks of COVID-19 disinformation to public health and safety. Countries across the globe have seen a spike in anti-Asian, anti-Semitic and other targeted hate, often directly citing or fuelled by conspiracies surrounding the virus' origin and transfer. At the same time, debunked theories related to 5G have spurred violent attacks against telecoms infrastructure and related personnel in the UK, Ireland, Belgium and the Netherlands. Conspiracy theories have not only sparked protests in the US, Australia, Germany and the UK (to cite just a few), but are helping promote scepticism and distrust in any future vaccine that might curb the virus' spread. If such trends continue, they will hinder any efforts to keep the public safe and well- informed.

This report offers an interim review of responses to the COVID-19 'infodemic' from three major technology companies - Facebook, Google and Twitter - from March to May 2020. These platforms have been forced to mobilise at speed, trialling policies and enforcement approaches that can meet such a challenge. The briefing summarises the approaches taken by respective teams at Twitter, Facebook, WhatsApp, Instagram, Google and YouTube, including specific services and policies introduced in recent months and, where possible, the accompanying rationale from companies themselves.

Such measures include:

- COVID-19 information hubs that share verified updates from sources like the World Health Organisation (WHO), Centres for Disease Control and Prevention (CDC) and national health ministries, including guidance tailored by geography;
- Partnerships with independent fact-checking networks such as Poynter IFCN to verify or debunk claims around the pandemic;
- Labelling, downranking and/or removing content flagged as false or misleading by experts;
- Official health alerts prompted by 'coronavirus' and related search terms;
- Prohibiting ads that aim to profiteer off the pandemic, including inflated prices for Personal Protective Equipment and unproven remedies, diagnostic tests or cures;
- Updated moderation policies to cover broadened definitions of 'harm', including content that contradicts public health guidance, creates panic based on fake claims, impersonates government officials, circulates unverified advice, and/or promotes scapegoating of certain groups;
- Free advertising credits for government and multilateral public health bodies, to increase the visibility of key guidance and updates.

To evaluate the success of these measures, the report combines evidence from ISD's own Digital Analysis Unit, which has published regular briefings on COVID-19 Disinformation, with recent data and research compiled from Avaaz, Media Matters, the BBC, The Telegraph, the New York Times, the Reuters Institute, the Oxford Internet Institute, ProPublica, Graphika, the University of Ottawa, Carleton University, Ottawa Hospital, the Tech Transparency Project, the Australia Institute Centre

for Responsible Technology, Queensland University of Technology, Consumer Reports, and others.
The collected evidence indicates that efforts have failed to stem the tide of disinformation, weaponised hate, profiteering, conspiracy theories and other harmful behaviours surrounding the pandemic. In particular, it finds a continual disconnect between the formulation and intent of new policies, and their comprehensive enforcement on and across platforms.

Analysis is clustered under three areas of platform policy - content moderation, advertising and proactive information – with case studies that highlight certain key flaws and challenges to combatting disinformation online. These include the following:

1) False and misleading content around COVID-19 is still widely circulated, despite being flagged by experts. The review finds that content debunked by fact-checkers, as well as websites hosting known mis- and disinformation around COVID-19, have been shared millions of times across social media platforms, often without labels or warnings. Moreover, the level of user engagement with known mis- and disinformation appears to dwarf that of parallel content from the WHO and other verified experts in many instances. This disparity exists despite platforms' efforts to promote verified information, including via alert boxes, knowledge panels and other push notifications.

2) Extremists are hijacking COVID-19 content to spread their message. Across the ideological spectrum, extreme groups are weaponising the pandemic to increase traffic and visibility for their cause online. This includes the co-option of relevant hashtags, avatars and trending topics by ISIS-linked networks, and the creation of so-called 'coronavirus' pages that funnel users to violent extremist content. Research indicates a spike in discussions around the 'boogaloo', a term used by the far-right to describe an impending 'second civil war', alongside public groups aimed to mobilise citizens for armed insurrection and targeted attacks (e.g. deliberately infecting politicians, journalists, front-line health providers, key workers and ethnic minorities). In parallel, Islamist groups badged as 'health and wellbeing' are celebrating the death toll in Anti-Daesh Coalition states and linking followers to ISIS media outlets like al-Naba, Muslim News, The Punishment and al-Bayan Radio.

3) Automated and inauthentic accounts are promoting COVID-19 disinformation and the related policy agendas of foreign states. Thousands of presumed inauthentic and sock-puppet accounts are being used to promote COVID-19 disinformation on Twitter and Facebook, including for explicit political gain. Tactics include coordinated bot, human-bot hybrid and fully-human co-retweet networks, some of which contain hacked profiles or those purchased as 'inactive'.

4) Google and Facebook continue to host advertisements banned under their new COVID-19 guidelines. Paid advertising is being used to profiteer off the pandemic and spread harmful messaging, despite apparent bans from platforms. Published posts include the sale of unverified therapies and 'medical-grade' equipment, as well as sponsored content that claims the virus is a hoax or promotes other related conspiracies. The delayed roll-out of policies around Political Ad Transparency has also enabled foreign states to run undisclosed advertising, including posts on Facebook and Instagram from Xinhua News Agency, Global China Television Network (GCTN) and China Central Television (CCTV) in English, Chinese and Arabic.

Sadly, any conclusions drawn must rely on some element of extrapolation and inference. Without better access to data and insight on companies' decision-making systems, both human- and machine-led, we cannot determine with certainty why some areas of policy appear more effective or better enforced than others. The disinformation incidents outlined in this report were exposed despite minimal data access - one can only imagine the real scale of the problem on those platforms, or what could be achieved with more candid partnerships between the tech and research sectors.

Recommendations: Building Systems Resilient to Disinformation

Coronavirus has been a sobering moment in the fight against disinformation, forcing tech companies to reassess whether their policies and enforcement are fit-for-purpose. Platforms have pivoted to an even greater reliance on AI systems to identify and classify harmful content; a move which partly reflects the scale of COVID-related activity, but primarily how the crisis has impacted moderation teams. With many teams now furloughed or sheltering in place, they have been unable to work remotely due to privacy and data security concerns.

At some point, the COVID-19 crisis will end or become a managed part of public health systems worldwide, and companies will resume a level of 'normal' business operations. It would be naive to assume the so-called 'infodemic' will follow suit, or that company systems will become more resilient on their own. We must learn from the acute challenges of this moment and the flaws it has exposed in the ability to prevent, identify and counter disinformation online.

Beyond issues of false content, the evidence available signals an urgent need to address the actors, behaviours and distribution mechanisms involved in disinformation. Such measures will be vital if companies are to mount a response proportional to the scale and nature of the current threat. The recommendations in this report chart a course for more effectively preventing and countering such activity, including:

- Robust transparency standards and oversight of the protocols that govern information flows, including the recommendation, curation and moderation systems used by platforms (both algorithmic and human);
- Reassessing the concepts that undergird platform responses to disinformation to prioritise identification of actors (state and non-state), behaviours (misrepresentative accounts or networks) and dissemination mechanisms involved, rather than the nature of content itself;
- More consistent enforcement of existing policies across issues and actors, including the nexus between targeted hate and disinformation;
- Measures to increase 'friction' for disinformation actors, including the insertion of automated 'break points' in rapid news spikes, allowing verification and human vetting before a story goes viral;
- Formalised cross-platform partnership to tackle information crises, involving both tech giants and smaller or emerging platforms, and drawing on existing models (e.g. those related to terrorist content and child sexual exploitation);
- Better engagement with platform influencers, who can act as 'credible messengers' in times of crisis and support the communication efforts of formal institutions. This could include reimagining the form and nature of official alerts, encouraging higher 'brand values' to command user attention, as well as training for influencers so they understand the most harmful mis- and disinformation trends circulating online.

Introduction: A new kind of information crisis

Since January 2020, COVID-19 has become the perfect crucible for online harms. Pandemics are by their nature fast-moving, with constantly evolving information even from credible and expert sources. This is set against a backdrop of heightened fear and anxiety, where valid concerns over resource scarcity, economic fallout and personal safety merge with extremist views on race and social order. New conspiracies and coordinated disinformation efforts have exploded online, preying on the uncertainty of this moment and the ambiguity regarding the origins and spread of the disease worldwide.

Tech platforms have been forced to mobilise at speed, trialling policies and enforcement approaches that can meet such a challenge. Initial moves suggested the crisis could prove a turning point, not only in how companies conceptualise and enact their obligations towards users, but concerning their role in wider public safety. To date, companies have largely rejected or underplayed any causal link between online content and offline harm, including explicit violence, and kept external parties at arm's length when crafting their moderation policies.

These dynamics have shifted during the pandemic, both due to mounting evidence on the correlation between online trends and offline behaviour, and because liabilities surrounding inaction were identified by platforms at an early stage. Companies have engaged experts to support their efforts, attempting to mitigate the real-world risks of disinformation and manipulation on their platforms. This includes work with health bodies such as the World Health Organisation (WHO), Centres for Disease Control and Prevention (CDC) and national ministries, alongside trusted academics and fact-checking NGOs, focussing on two key goals: first, to surface authoritative health information around the pandemic, and second, to mitigate the spread of false or misleading content.

Measures include:

- Information hubs that share updates from verified sources, often tailored by geography;
- Partnerships with independent fact-checking networks to verify or debunk claims around the pandemic;
- Labelling, downranking and/or removing content flagged as false or misleading by fact-checkers;
- Official alerts prompted by coronavirus and related search terms;
- Prohibiting ads that aim to profiteer off the pandemic, including inflated prices for Personal Protective Equipment and unproven remedies, diagnostic tests or cures;
- Updated moderation policies to cover broadened definitions of 'harm', including content that contradicts public health guidance, creates panic based on fake claims, impersonates government officials, circulates unverified advice, and/or promotes scapegoating of certain groups;
- Free advertising credits for government and multilateral public health bodies.

While this is undoubtedly progress, the limits of reactive efforts have become clear as the situation unfolds. Despite emergency measures, the systems for decision-making and policy application have been tested and proven themselves fatally constrained; there remain inherent flaws that limit a more comprehensive or scaled approach to prevent, mitigate and counter disinformation.

While there are immediate steps which can and should be taken to stem harmful trends, the threat expands beyond coronavirus and demands a broader reflection on our digital landscape. This crisis will not be the last, whether in the public health sphere or surrounding other global challenges like climate change, natural disasters and inter-state conflict. It is therefore crucial to learn from the current pandemic, and build the architecture for a more systemic response going forward. By analysing key gaps in the systems that govern major platforms, we can develop a long-term blueprint for crisis management and build systems that are more resilient to disinformation efforts writ large.

For the past decade, ISD has been analysing and responding to a range of online harms, including violent extremism, disinformation and hate speech. Since COVID-19 emerged, we have tracked how the public health crisis is exacerbating these threats and the respective attempts at prevention and mitigation from social media companies. For this review we have assessed efforts to deal with disinformation in particular, citing evidence around three domains of policy design and enforcement:

- Content moderation
- Advertising
- Proactive information

There is a growing body of research that analyses, probes and sense-checks company practices around advertising and content moderation - some key case studies are compiled in Section 2, alongside a wider range of data and reporting which has informed our conclusions. There is currently no equivalent, publicly-available data with which to assess proactive information policies, including the 'alert boxes' and 'information hubs' that are increasingly used on platforms like Facebook, Facebook Messenger, Twitter and YouTube. Measures to promote evidence-based and verifiable information to users is key, especially in times of crisis, but without impact data it is difficult to gauge their effectiveness in practice.

This paper is an interim review of responses since February 2020, focusing on three major social media platforms - Google[1] (-including YouTube), Facebook[2] (including WhatsApp and Instagram) and Twitter[3]. It acknowledges the efforts made by companies, which are summarised in Annex 1, while also providing suggestions for how efforts to prevent or counter disinformation could be strengthened and implemented. We have also included a timeline of when actions were announced or applied by platforms, mapped onto developments in the spread of the pandemic and specific offline events that prompted a response. The assessment below is by no means all-encompassing, but provides a snapshot of publicly available evidence to guide further action.

Overall, the 'infodemic' surrounding COVID-19 has exposed a digital system that is underprepared to tackle viral lies and the misrepresentation of identity, attention, and popularity.

Beyond issues of false content, this signals the urgent need to address actors, behaviours and distribution mechanisms involved in disinformation. Such measures will be vital if companies are to mount a response proportional to the scale and nature of the current threat. The recommendations in this report chart a course for more effectively preventing and countering disinformation.

[1] Coronavirus disease 2019 (COVID-19) updates. Youtube Help. https://support.google.com/youtube/answer/9777243?hl=en

[2] Jin, K., Keeping People Safe and Informed About the Coronavirus. Facebook. https://about.fb.com/news/2020/06/coronavirus/

[3] Coronavirus: Staying safe and informed on Twitter. Twitter. https://blog.twitter.com/en_us/topics/company/2020/covid-19.html

Online Content, Offline Harm: Why Disinformation Matters

The disinformation crisis surrounding COVID-19 is not an abstract problem. Online content can help catalyse real-world harm, and research is already documenting the risks of COVID-19 disinformation to both public health and safety. An April 2020 study from King's College London[4] showed a statistical link between belief in three prominent conspiracy theories around coronavirus and non-compliance with related public health guidelines. There has been a documented increase in anti-Asian hate crime across many countries, including the US[5] and UK[6], accompanying unfounded claims online that COVID-19 was released from a Wuhan laboratory or orchestrated by the Chinese State. Countries have witnessed a rise in conspiracy-fuelled violence against 5G property and related individuals during the pandemic, including attacks against critical telecoms infrastructure[7] and personnel in the UK, Ireland, Belgium and the Netherlands[8]. Meanwhile, content claiming the pandemic was planned by so-called 'elites'[9] threatens to erode trust in institutions attempting to mitigate the public health crisis. These conspiracy theories have not only fuelled protests in the US[10], Australia[11], Germany[12] and the UK[13] (to cite just a few), but are helping promote scepticism[14] and distrust in any future vaccine that might curb the virus' spread. If such trends continue, they will hinder any efforts to keep the public safe and well- informed. Further evidence and analysis can be found in ISD's COVID-19 Disinformation Briefings, which are published on a regular basis by the Digital Analysis Unit.

Protecting the integrity of how people select and receive information is of utmost importance during times of crisis. The recommendations in this report reflect the urgent need for transparency and accountability regimes, unpacking how companies' design choices and information control systems impact their user base. Oversight for such a process will vary from context to context: democratic governments may be well-placed to design and spearhead efforts, whilst elsewhere independent experts or even voluntary action from companies remain the only viable options. Nonetheless, the same principles must undergird any future efforts to confront disinformation online. The case studies detailed below offer a compelling rationale for change, alongside the vast array of research which could not be captured here - the status quo is failing to deliver at scale, and COVID-19 provides an optimum moment for reform.

[4] Allington, D., & Dhavan, N. (2020). The relationship between conspiracy beliefs and compliance with public health guidance with regard to COVID-19. Centre for Countering Digital Hate. https://kclpure.kcl.ac.uk/portal/files/127048253/Allington_and_Dhavan_2020.pdf

[5] Zhou, L. (2020, April 21). How the coronavirus is surfacing America's deep-seated anti-Asian biases. Vox. https://www.vox.com/identities/2020/4/21/21221007/anti-asian-racism-coronavirus

[6] Grierson, J. (2020, May 13). Anti-Asian hate crimes up 21% in UK during coronavirus crisis. Guardian. https://www.theguardian.com/world/2020/may/13/anti-asian-hate-crimes-up-21-in-uk-during-coronavirus-crisis

[7] Temperton, J. (2020, April 6). How the 5G coronavirus conspiracy theory tore through the internet. Wired. https://www.wired.co.uk/article/5g-coronavirus-conspiracy-theory

[8] Chan, K, Dupuy, B & Lajka A. (2020, April 21). Conspiracy theorists burn 5G towers claiming link to virus. CTV News. https://www.ctvnews.ca/health/coronavirus/conspiracy-theorists-burn-5g-towers-claiming-link-to-virus-1.4905039

[9] Funke, D. (2020, May 7). Fact-checking 'Plandemic': A documentary full of false conspiracy theories about the coronavirus. PolitiFact (The Poynter Institute). https://www.politifact.com/article/2020/may/08/fact-checking-plandemic-documentary-full-false-con/

[10] Bogel-Burroughs, N. (2020, May 2). Antivaccination Activists Are Growing Force at Virus Protests. New York Times. https://www.nytimes.com/2020/05/02/us/anti-vaxxers-coronavirus-protests.html

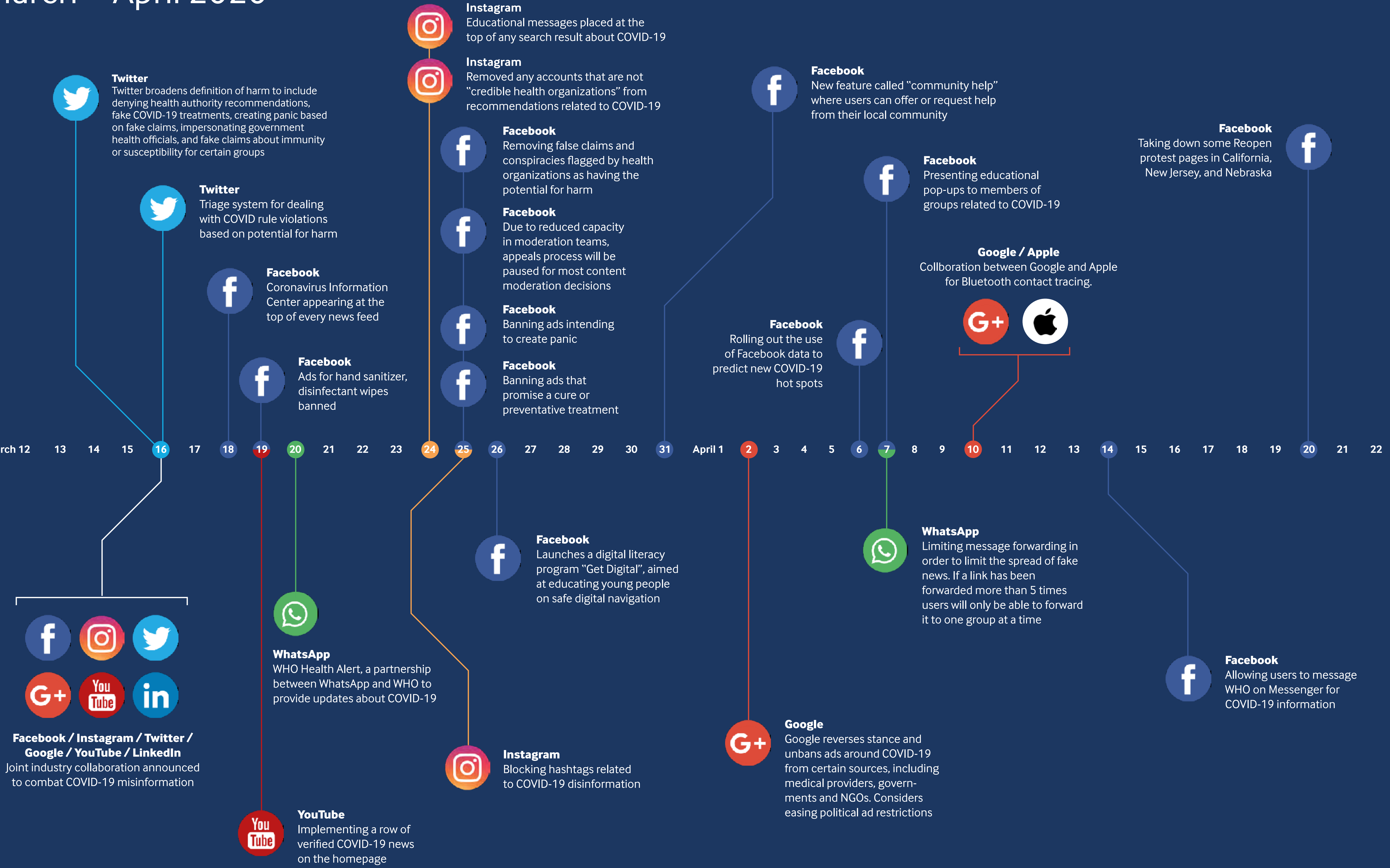
[11] Wilson, C. (2020, May 11). Why Are Australians Chanting "Arrest Bill Gates" At Protests? This Wild Facebook Group Has The Answers. Buzzfeed News. https://www.buzzfeed.com/cameronwilson/lockdown-protest-australia-bill-gates-conspiracy-theories

[12] Ankel, S. (2020, May 24). Germany is at the forefront of a global movement of anti-vaxxers obsessed with Bill Gates and it could mean the coronavirus is never defeated. Business Insider. https://www.businessinsider.com/germany-becomes-forefront-of-a-global-movement-of-anti-vaxxers-2020-5?r=US&IR=T

[13] Dearden, L. (2020, May 16). Coronavirus: Inside the UK's biggest anti-lockdown protest. The Independent. https://www.independent.co.uk/news/uk/home-news/coronavirus-lockdown-protests-uk-london-hyde-park-5g-conspiracy-theories-a9518506.html

[14] Wilson, C. (2020, May 20). As The World Hopes For A COVID-19 Vaccine, Anti-Vaxxers Are Growing Their Social Media Influence. Buzzfeed News. https://www.buzzfeed.com/cameronwilson/coronvirus-antivaxxers-facebook-instagram-boost

Timeline of Platform Responses: March – April 2020



Twitter
Twitter broadens definition of harm to include denying health authority recommendations, fake COVID-19 treatments, creating panic based on fake claims, impersonating government health officials, and fake claims about immunity or susceptibility for certain groups

Twitter
Triage system for dealing with COVID rule violations based on potential for harm

Facebook / Instagram / Twitter / Google / YouTube / LinkedIn
Joint industry collaboration announced to combat COVID-19 misinformation

Facebook
Coronavirus Information Center appearing at the top of every news feed

Facebook
Ads for hand sanitizer, disinfectant wipes banned

YouTube
Implementing a row of verified COVID-19 news on the homepage

WhatsApp
WHO Health Alert, a partnership between WhatsApp and WHO to provide updates about COVID-19

Instagram
Educational messages placed at the top of any search result about COVID-19

Instagram
Removed any accounts that are not "credible health organizations" from recommendations related to COVID-19

Facebook
Removing false claims and conspiracies flagged by health organizations as having the potential for harm

Facebook
Due to reduced capacity in moderation teams, appeals process will be paused for most content moderation decisions

Facebook
Banning ads intending to create panic

Facebook
Banning ads that promise a cure or preventative treatment

Instagram
Blocking hashtags related to COVID-19 disinformation

Facebook
Launches a digital literacy program "Get Digital", aimed at educating young people on safe digital navigation

Facebook
New feature called "community help" where users can offer or request help from their local community

Google
Google reverses stance and unbans ads around COVID-19 from certain sources, including medical providers, governments and NGOs. Considers easing political ad restrictions

Facebook
Rolling out the use of Facebook data to predict new COVID-19 hot spots

Facebook
Presenting educational pop-ups to members of groups related to COVID-19

WhatsApp
Limiting message forwarding in order to limit the spread of fake news. If a link has been forwarded more than 5 times users will only be able to forward it to one group at a time

Google / Apple
Collboration between Google and Apple for Bluetooth contact tracing.

Facebook
Allowing users to message WHO on Messenger for COVID-19 information

Facebook
Taking down some Reopen protest pages in California, New Jersey, and Nebraska

Evaluation: Assessing Companies' Policies and Actions around COVID-19

1. Content Moderation

Where was the policy position previously?

Over the past five years, the approaches taken by Twitter, Facebook and Google in addressing disinformation have differed significantly in both definition and detail, yet the overarching concepts are fairly consistent. **Each company has acknowledged the potential risks of false information on their platforms and subsequently amended Terms of Service or Community Guidelines with new policies**. This has primarily centred on efforts to label content identified as false or misleading by independent fact-checkers, sometimes accompanied by measures to curb the spread of such content and/or users' ability to engage with it on sites like Twitter and Facebook. YouTube has used existing online resources such as Wikipedia and Encyclopedia Britannica as information cues on conspiracy theories deemed to be false or misleading.[15] Companies have largely relied on a handful of independent experts to identify and verify claims, a process that swiftly revealed a tension between the scale of online content and proportional, manual fact-checking resources.

Companies have also recognised how the misrepresentation of individuals, communities or popularity on their platforms can fuel disinformation online. Tackling the burgeoning field of tactics and actors required a wholly different approach for both detection and response, beyond those developed for false content alone. Overwhelmingly, attempts to combat disinformation have relied on a mix of in-house detection mechanisms (largely opaque to the outside world), and reactive measures to specific user or expert reports, flagging content, accounts or channels deemed to violate company Terms of Service.

Across the board, **companies have struggled with questions surrounding the application of such policies across different types of user**; this includes recent divergence[16] over whether to firewall public figures, especially politicians and leaders, who violate disinformation policies online. Facebook has gone further to protect certain kinds of actors in this respect, refusing fact-checking on adverts from politicians outright,[17] and publicly arguing that the platform should not be an 'arbiter of truth'. Such decisions have become only more controversial amid a global public health crisis.

What has changed because of coronavirus?

Companies have engaged with the outside world to hone definitions of 'harm' and 'fact', after a historic reluctance to engage independent bodies in policy formulation. TikTok has committed to forming a "committee of experts" to advise on moderation policies; Twitter issued a "broadened definition of harm", whereby tweets can be removed for denying health authority recommendations; Facebook have launched a 24-hour response team with health authorities to stay abreast of evidence and false claims. While in the past platforms were hesitant to involve external figures or groups in moderation, they are now at pains to align with the World Health Organisation (WHO), Centres for Disease Control and Prevention (CDC) and other national ministries, as well as their longer-running partnerships with fact-checking bodies like Poynter IFCN.[18]

[15] Matsakis, L. (2018, March 13). YouTube Will Link Directly to Wikipedia to Fight Conspiracy Theories. Wired. https://www.wired.com/story/youtube-will-link-directly-to-wikipedia-to-fight-conspiracies/

[16] Ghaffary, S. (2020, May 29). Facebook and Twitter have similar policies. But only Twitter is fighting Trump. Vox. https://www.vox.com/recode/2020/5/29/21275173/twitter-facebook-trump-executive-order-fact-check-freedom-of-speech-censorship-google

[17] https://about.fb.com/news/2019/09/elections-and-political-speech/

[18] https://www.poynter.org/ifcn/

Broadened definitions of harm include content that contravenes official health advice (e.g. on prevention, remedies, diagnosis and lockdown measures), or that aims to generate panic based on false claims (e.g. around food shortages, the immunity of certain groups). Twitter has introduced a special disclaimer[19] for posts containing COVID-19 misinformation, responding to such content under three categories: misleading information (labelled or removed), disputed claims (labelled and warning) and unverified claims (no action, unless it contravenes the new Content Policy). Facebook have incorporated false claims and conspiracies flagged by health bodies under their Content Violation measures, and partnered with 60 fact-checking organisations operating in 50 languages worldwide. Instagram claims to have removed COVID-19 accounts from the platform's 'Recommendations' function, unless linked to a known health agency, and content is intended to be down-ranked in Feed and Stories if rated false by third-party fact-checkers.

Under such guidelines, **companies have begun to enforce removal[20] or fact-checking on content from public figures**, including posts from Presidents Bolsonaro[21] and Maduro[22], and former New York Mayor and attorney to President Trump, Rudy Giuliani[23]. According to both Twitter and Facebook, the content violated their respective Community Guidelines for COVID-19, either by contradicting public health advice or promoting unproven remedies like hydroxychloroquine. That said, **crackdowns on misinformation from political leaders are still isolated and sporadic, with inconsistent application across different countries, issues, and platforms**.

Enforcement processes for these policies have also shifted under the weight of the crisis. **Platforms have pivoted to an even greater reliance on AI systems to identify and classify harmful content**; a move which partly reflects the scale of COVID-related activity, but primarily how the crisis has impacted moderation teams, who are often outsourced to external contractors rather than employed by companies themselves. With many teams now furloughed or sheltering in place, they have been unable to work remotely due to privacy and data security concerns. Such changes impact the detection and processing of violating content, but also limit the potential for redress in cases of wrongful removal or sanctions against a user. Facebook has announced it will pause moderation appeals for most content due to reduced staff capacity, while Twitter has noted the difficulty in taking 'enforcement action' on every Tweet containing incomplete or disputed information about COVID-19. There is an alarming lack of transparency around AI systems in content moderation, including the potential impact on the frequency of false negatives and false positives. In an effort to spur relevant research efforts during the crisis, ISD joined 75 organisations and experts in an open letter to tech companies[24]; the statement calls on social media and content-sharing platforms to preserve automated moderation data, making it available to researchers and journalists, and to include such data in future transparency reports.

For a detailed summary of each platform's response, please refer to **Annex 1**.

[19] Roth, Y & Pickles, N. (2020, May 11). Updating our Approach to Misleading Information. Twitter blog. https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html

[20] Lyons, K. (2020, March 30). Twitter removes tweets by Brazil, Venezuela presidents for violating COVID-19 content rules. The Verge. https://bit.ly/2YsfYr1

[21] Wagner, K. (2020, March 31). Facebook, Twitter, YouTube Remove Posts From Bolsonaro. Bloomberg. https://www.bloomberg.com/news/articles/2020-03-31/facebook-twitter-pull-misleading-posts-from-brazil-s-bolsonaro

[22] Coronavirus: World leaders' posts deleted over fake news (2020, March 31). BBC News. https://www.bbc.co.uk/news/technology-52106321

[23] Porter, T. (2020, March 29). Twitter deleted a tweet by Rudy Giuliani for spreading coronavirus misinformation. Business Insider. https://www.businessinsider.com/coronavirus-twitter-deletes-giuliani-tweet-for-spreading-misinformation-2020-3?r=US&IR=T

[24] COVID-19 Content Moderation Research Letter – in English, Spanish, & Arabic. (2020, April 22). Centre for Democracy & Technology. https://cdt.org/insights/covid-19-content-moderation-research-letter/

Content Moderation **Case Study 1**

Misinformation and conspiracy theories receiving tens of millions of views on Facebook

ISD[25], Avaaz[26] and Media Matters[27]

Researchers continue to expose widespread violations of Facebook's Terms of Service around COVID-19, with the often unchecked spread of health mis- and disinformation across the platform. This includes the presence of Facebook groups and pages dedicated to promoting false and misleading claims around the pandemic.

Campaign organisation Avaaz[28] examined over 100 pieces of misinformation content about coronavirus in six languages, which had already been rated false or misleading by reputable, independent fact-checkers. The research found that 'pieces of content ... sampled and analysed were shared over 1.7 million times on Facebook, and viewed an estimated 117 million times'.

100+
pieces of flagged disinformation shared over **1.7m times** on Facebook with **117m views** (Avaaz)

Media Matters[29] identified ten companies promoting misinformation about coronavirus cures and prevention via Facebook pages, amassing hundreds of thousands of followers. The organisation also reported on the residual presence of 'ReOpen' Facebook events and groups[30] violating Stay-at-home orders, despite policy announcements made by the platform that it had removed such content.

A recent investigation by ISD and the BBC[31] found that websites known to host disinformation about Coronavirus had received over 80 million interactions on public Facebook pages since the start of the year. As a benchmark, in the same period links to the CDC and WHO websites gathered around 12 million interactions combined. These websites were shared in public groups and pages on Facebook that also included extremist and hateful content about coronavirus, including posts linking Jews and Muslims to the creation of the coronavirus or its spread. Facebook removed a number of the posts referenced in the research having been notified by the BBC.

A New York Times analysis[32] found posts pushing unproven UV light therapies for the coronavirus in 780 Facebook groups, 290 Facebook pages, 9 Instagram accounts and 'thousands' of tweets.

WHAT IS THE IMPLICATION?

The sheer scale of unchecked mis- and disinformation on Facebook renders post-hoc fact checking of limited use, and will not prove sufficient to stem the tide of false health content. This evidence demonstrates how vital it is to interrogate content promotion and amplification systems, which enable the rapid and uncontrolled spread of false and misleading content. It also indicates the need for 'stage-gates' to prevent disinformation from reaching a tipping point of engagement on platforms, before they are detected and blocked (See: Recommendation One).

[25] https://www.bbc.co.uk/iplayer/episode/m000hzd6/click-a-changing-world

[26] Avaaz. (2020). How Facebook can Flatten the Curve of the Coronavirus Infodemic. https://secure.avaaz.org/campaign/en/facebook_coronavirus_misinformation/

[27] Hananoki, E. (2020, May 1). Facebook says it's removing content promoting false coronavirus preventatives and cures. These businesses are currently violating that policy. Media Matters. https://www.mediamatters.org/coronavirus-covid-19/facebook-says-its-removing-content-promoting-false-coronavirus-preventatives

[28] Avaaz. (2020). How Facebook can Flatten the Curve of the Coronavirus Infodemic. https://secure.avaaz.org/campaign/en/facebook_coronavirus_misinformation/

[29] Hananoki, E. (2020, May 1). Facebook says it's removing content promoting false coronavirus preventatives and cures. These businesses are currently violating that policy. Media Matters. https://www.mediamatters.org/coronavirus-covid-19/facebook-says-its-removing-content-promoting-false-coronavirus-preventatives

[30] Gogarty, K. (2020, April 4). Facebook says it removed events violating stay-at-home orders. But, it hasn't removed them. Media Matters. https://www.mediamatters.org/coronavirus-covid-19/facebook-says-it-removed-events-violating-stay-home-orders-it-hasnt-removed

[31] https://www.bbc.co.uk/iplayer/episode/m000hzd6/click-a-changing-world

[32] Frenkel, S. & Alba, D. (2020, April 30). Trump's Disinfectant Talk Trips Up Sites' Vows Against Misinformation. New York Times. https://www.nytimes.com/2020/04/30/technology/trump-coronavirus-social-media.html?smtyp=cur&smid=tw-nytimes

Content Moderation **Case Study 2**

Inauthentic account networks creating and promoting COVID-19 material on Twitter

ProPublica[33], Graphika[34], and The Australia Institute (Centre for Responsible Technology) and Queensland University of Technology (Digital Media Research Centre)[35]

Beyond dealing with false content outright, companies are failing to address actors that conduct covert platform manipulation, including the infiltration and misrepresentation of communities at scale. The range of perpetrators include those linked to foreign states, as well as many non-state domestic and transnational groups.

PBeyond dealing with false content outright, companies are failing to address actors that conduct covert platform manipulation, including the infiltration and misrepresentation of communities at scale. The range of perpetrators include those linked to foreign states, as well as many non-state domestic and transnational groups.

Since August 2019, ProPublica[36] has tracked over 10,000 suspected fake Twitter accounts stated to be part of a 'coordinated influence campaign with ties to the Chinese government'. Those accounts, many of which appear to be hacked from users around the world or purchased as old/inactive profiles, are being co-opted to promote 'propaganda and disinformation about the coronavirus outbreak, the Hong Kong protests and other topics of state interest'.

Researchers from Graphika[37] analysed Iranian state-supporter networks on Twitter, specifically the International Union of Virtual Media (IUVM) network of websites and social media assets. Graphika explored how these networks were used to spread disinformation about COVID-19, typically involving the creation or copying of news, cartoons or videos to promote Iranian government messaging on social media. Some accounts were deemed to be posing as journalists or official sources, and Twitter removed one account flagged to them through the research. Related networks were also analysed on Facebook and Instagram.

The Australia Institute (Centre for Responsible Technology) and Queensland University of Technology (Digital Media Research Centre) analysed 2.6 million tweets relating to coronavirus and their 25.5 million retweets over a 10-day period, focusing on one conspiracy relating to the pandemic: that China bioengineered the virus as a weapon, and it was either accidentally or strategically released from a virology lab in Wuhan. From this dataset they identified 5,752 accounts that coordinated 6,559 times to spread mis- and disinformation regarding COVID-19, for either commercial or political motives.

The teams conducted two forms of co-retweet analysis: the first using a 'bot' threshold (co-retweets <1 second) and the second a Keller approach for inauthentic behaviour (frequently co-retweeting <1 minute). This helped to capture fully automated bot accounts, hybrid automated-human accounts and fully human accounts

[33] Kao, J. & Li, M.S. (2020, March 26). How China Built a Twitter Propaganda Machine Then Let It Loose on Coronavirus. ProRepublica. https://www.propublica.org/article/how-china-built-a-twitter-propaganda-machine-then-let-it-loose-on-coronavirus

[34] Nimmo, B. et al. (2020, April 15). Iran's IUVM Turns To Coronavirus. Graphika. https://graphika.com/reports/irans-iuvm-turns-to-coronavirus/

[35] Graham, T. et al. (2020, May). Like a virus The coordinated spread of coronavirus disinformation. https://bit.ly/3fcN9Wb

[36] Kao, J. & Li, M.S. (2020, March 26). How China Built a Twitter Propaganda Machine Then Let It Loose on Coronavirus. ProRepublica. https://www.propublica.org/article/how-china-built-a-twitter-propaganda-machine-then-let-it-loose-on-coronavirus

[37] Nimmo, B. et al. (2020, April 15). Iran's IUVM Turns To Coronavirus. Graphika. https://graphika.com/reports/irans-iuvm-turns-to-coronavirus/

working in tandem. 30 clusters emerged from the retweet network, 28 of which were pro-Trump, actively promoting the QAnon conspiracy theory and/or hyper-partisan Republican positions. Co-ordinated efforts focussed on 882 original tweets, retweeted 18,498 times and liked 31,783 times, with an estimated 5 million impressions overall. During the study, the team also identified multiple networks coordinating to exploit the 'trending topics' algorithms on Twitter.

Some bot networks were incidental (i.e. not explicitly malicious in intent), such as accounts retweeting prizes and giveaway competitions, creating spam for comedic purposes, promoting company services or products, or part of a positive campaign (e.g. '100 days of coding', a campaign promoting AI to understand the spread of COVID-19 and reposting news articles about IT security). However ten bot-like networks were clearly orchestrated for disinformation, including use of hacked accounts, co-retweeting and astroturfing methods. Notable examples seemed designed to:

Magnify fear and sow discord about the COVID-19 mortality and infection rates in Paraguay;

- Aggravate political tension re: the government response and curfew restriction in Turkey (including opposition between President Erdoğan and Istanbul's more moderate opposition mayor, Ekrem Imamoglu);
- Blanket the site with positive endorsements of Saudi Crown Prince Mohammed bin-Salman;
- Amplify political tensions in Spain by retweeting hyper-partisan content (e.g. criticism of the official response, memes portraying the government as fascists).

WHAT IS THE IMPLICATION?

Companies must expand efforts to detect and remove inauthentic networks of accounts, including those who seek to profit from or exploit crisis situations, no matter which actors are responsible. To address the current threat, platforms need to widen their detection efforts to ensure they are capturing manipulation enacted by non-state actors and groups as much as foreign states, each of which pose different kinds of challenges to civil rights, public health and democratic processes (See: Recommendation Two).

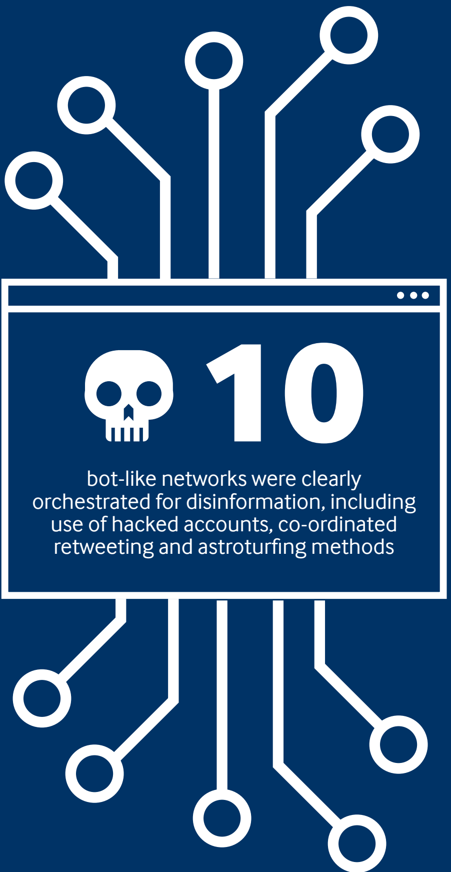
AT A GLANCE

10,000+

uspected inauthentic Twitter accounts

2.6m

Number of tweets analysed by the Australia Institute and Queensland University of Technology



10

bot-like networks were clearly orchestrated for disinformation, including use of hacked accounts, co-ordinated retweeting and astroturfing methods

Content Moderation **Case Study 3**

YouTube pushing misleading or false information through video search results on Coronavirus

University of Ottawa, Carleton University and Ottawa Hospital[38]

Accurate and evidence-based video content is systematically under-represented in search results on YouTube, both in regard to COVID-19 and other health crises.

According to a March study published in BMJ Global Health[39], over 25% of the most viewed YouTube videos about coronavirus contain false or misleading information. Researchers used blanket searches with the keywords 'coronavirus' and 'COVID-19' and analysed the top 75 results in each instance. Videos that were duplicates, non-English language, lacking audio/visuals, longer than 1 hour, livestreams or unrelated to the pandemic were excluded, leaving 69 pieces of content with nearly 258 million views at the point of search. Of these 69, over a quarter (27.5%) contained non-factual information, with an aggregated reach of over 62 million views. The videos were categorised under seven headings, but most views fell under Network News, Consumers and Entertainment News (72% combined).

WHAT IS THE IMPLICATION?

YouTube has been criticised for allowing misinformation to thrive, including in previous emergencies such as the H1N1 pandemic[40], and Ebola[41] and Zika[42] outbreaks. In all cases, a similar proportion of videos (around 25%) were found to be misleading in relation to the crisis, while reputable content was less visible; this includes Professional and Government videos which had a universally higher accuracy, usability and quality rating across all metrics in this study. YouTube must take greater pains to support public health bodies on engagement, making guidance less static and more aligned with viral content. This could include partnerships with content producers, or circulation via key influencers and channels. More importantly, greater oversight is required to assess the outcomes of YouTube's algorithmic decision-making on search results and video recommendations. In doing so, independent experts could track and quantify the role YouTube plays during public health crises, especially the promotion of disinformation over verified health content (See: Recommendation One).

[38] Li HO-Y, Bailey A, Huynh D, et al. (2020, April 24). YouTube as a source of information on COVID-19: a pandemic of misinformation? BMJ Global Health. https://gh.bmj.com/content/bmjgh/5/5/e002604.full.pdf

[39] ibid.

[40] Pandey A. et al. (2010, March 1). YouTube As a Source of Information on the H1N1 Influenza Pandemic. American Journal of Preventive Medicine. https://www.ajpmonline.org/article/S0749-3797(09)00806-X/fulltext

[41] Pathak, R., Poudel, D R., Karmacharya P. et al. (2015). Youtube as a source of information on Ebola virus disease. North American Journal of Medical Science, 7(7), 306-9. http://www.najms.org/article.asp?issn=1947-2714;year=2015;volume=7;issue=7;spage=306;epage=309;aulast=Pathak

[42] Bora, K., Das, D., Barman, B. & Probodh Borah (2018). Are internet videos useful sources of information during global public health emergencies? A case study of YouTube videos during the 2015–16 Zika virus pandemic. Pathogens and Global Health, 112(6), 320-328, https://www.tandfonline.com/doi/full/10.1080/20477724.2018.1507784

25%

of the most viewed YouTube videos about coronavirus contain false or misleading information

62m

Aggregated reach OF non-factual information

Content Moderation **Case Study 4**

Inconsistent use of fact-checking and warning labels on known false or misleading content

Reuters Institute for the Study of Journalism and the Oxford Internet Institute[43]

Efforts have been made to provide evidence-based information to users, particularly those who view or engage with information rated false by independent fact-checking organisations. Unfortunately, companies' enforcement of these policies is patchy at best, and does not proportionally meet the scale of disinformation circulating on their platforms.

Joint research[44] by the Reuters Institute for the Study of Journalism and the Oxford Internet Institute found that platforms varied significantly in their response to content deemed false or misleading by independent fact-checkers, with Twitter failing to take action on over half of the debunked content on its platform. The research, using lists of fact-checks from organisations such as First Draft, found that Twitter did not remove, label or otherwise action 59% of posts rated false in the sample; comparatively, 27% of content remained live on YouTube, and 24% on Facebook.

More recently, Twitter has updated its labelling policy[45] to include a broader range of content, including 'potentially harmful, misleading information related to COVID-19'. The new policy deals with three categories: misleading information, disputed claims, and unverified claims. No action will be taken on the latter, but labels, warnings and removals can be used for the former two groups. However, the systems used to identify and categorise content remain largely opaque, and it is unclear if the company can improve on past performance with these tools. Even since the new announcements, experts at the University of California Berkeley[46] have raised concerns about the mislabelling of 5G-related content through automated content identification on Twitter.

WHAT IS THE IMPLICATION?

The scale of debunked disinformation that is still viewable on major platforms without labelling suggests that new measures are needed. Even when fact-checkers debunk a claim on social media, the speed at which content is allowed to travel through and between platforms hinders any meaningful response; by the detection stage, there are often thousands of examples beyond the original piece of content, which would each need to be detected, removed or flagged. The question is whether instantaneous information should come at the expense of safety. Disinformation can travel across platforms and replicate for hours, if not days, gradually building a critical mass - if caught early enough in the trajectory, it would prove more feasible to add effective disclaimers. It is vital to introduce greater 'friction' as a topic begins to spike, enabling fact-checkers to review claims before they reach widespread visibility. This could include the insertion of 'breaking points' in rapidly emerging stories, activated if a piece of content starts to snowball within a given timeframe and at sufficient scale. (See Recommendation One). In addition, platforms should work collaboratively where possible to counter the spread of known disinformation material, building off similar models in other areas of online harm (e.g. terrorist material and images of child sexual exploitation). (See Recommendation Five).

[43] Brennen, J. S., Simon, F., Howard P. N. & Nielson, R. K. (2020, April 7). Types, sources, and claims of COVID-19 misinformation. Reuters. https://reutersinstitute.politics.ox.ac.uk/types-sources-and-claims-covid-19-misinformation#menu

[44] ibid.

[45] Roth, Y & Pickles, N. (2020, May 11). Updating our Approach to Misleading Information. Twitter blog. https://blog.twitter.com/en_us/topics/product/2020/updating-our-approach-to-misleading-information.html

[46] Wong, Q. (2020, May 25). More harm than good? Twitter struggles to label misleading COVID-19 tweets. Cnet. https://www.cnet.com/news/more-harm-than-good-twitter-struggles-to-label-misleading-covid-19-tweets/

Content Moderation **Case Study 5**

Coronavirus content used to promote hate and extremism on platforms

Tech Transparency Project[47] and ISD[48]

Extremists have used a variety of online platforms to spread disinformation[49] on COVID-19 transmission and undermine evidence-based policymaking or public health guidelines. In tandem, these communication spaces are rife with extremist and violent content, from documents preparing recruits for the 'second civil war' to calls for violence against Muslims, Jews, and frontline services like the police. Evidence has been found surrounding both Islamist and far-right account networks and pages, co-opting conversations about COVID-19 to spread extremist material online.

Tech Transparency Project's report[50] outlines how extremists are "using Facebook to organize for civil war amid coronavirus", including the policy violations of over a hundred Facebook groups mobilising for a violent insurrection and attracting 'tens of thousands of members' amid the crisis. Over 60% of the groups were created since the beginning of the year, as the pandemic took hold in the U.S. and other countries worldwide. Content in these groups involves discussion of 'tactical strategies, combat medicine, and various types of weapons, including how to develop explosives and the merits of using flamethrowers'.

ISD also reported on the prevalence of 'second civil war' extremists on Facebook in its second briefing[51] on COVID-19 disinformation. The report highlighted the communications network used by these groups to organise and share materials, often via messaging app Telegram as well as larger social media platforms. The research exposed over 200,000 posts containing the word 'boogaloo' across platforms (a term used for the impending rebellion), with 52% on Twitter, 22% on Reddit, 12% on Tumblr and 11% on 4Chan and Voat between 1 February and 28 March 2020. The research also highlighted how public Facebook groups are being used to discuss and mobilise around a second civil war, with posts ranging 'from satirical memes with violent implications to more explicitly extremist content'. These groups had between 6,500 and 22,000 followers, receiving 'significant increases in engagement over March' as coronavirus became a more prevalent topic of conversation. The admin posts alone from one group received 127,089 interactions in March 2020.

ISD research has revealed how networks of hijacked, hacked and repurposed accounts are co-opting COVID-19 topics on Facebook and Twitter to spread pro-ISIS messaging. ISD researchers tracked 230 ISIS-linked accounts on Twitter over a 30-day period between March and April, monitoring their use of coronavirus hashtags, avatars and themed videos to further their Arabic-language propaganda. Accounts targeted trending COVID-19 hashtags in Saudi Arabia, Lebanon, Jordan, Egypt, Iraq, Afghanistan and Pakistan, using multiple languages to create an ISIS "brand" in key regions. In one instance, an account named "Coronavirus" retweeted a pandemic-themed ISIS video to its thousands of supporters. Combined, the accounts generated more than 570,000 views of 155 pieces of official ISIS video content from Iraq, Syria, Afghanistan Somalia, Mozambique, and Nigeria, all piggybacking on COVID-19 hashtags. 38 percent of the tracked accounts linked to key ISIS websites, media aggregators and repositories, including 'Muslim News,' 'The Punishment,' and 'al Bayan Radio'. ISD also found the strategic use of Twitter Ads to spread ISIS content and attempt to drown out other COVID-19 related posts.
On Facebook, the Arabic language 'Coronavirus' page

[47] Tech Transparency Project (2020). Extremists Are Using Facebook to Organize for Civil War Amid Coronavirus. https://www.techtransparencyproject.org/articles/extremists-are-using-facebook-to-organize-for-civil-war-amid-coronavirus

[48] ISD (2020). Covid-19 Disinformation Briefing No.2. https://www.isdglobal.org/isd-publications/covid-19-disinformation-briefing-no-2/

[49] MacFarquhar, N. (2020, May 3). The Coronavirus Becomes a Battle Cry for U.S. Extremists. New York Times. https://www.nytimes.com/2020/05/03/us/coronavirus-extremists.html

[50] Tech Transparency Project (2020). Extremists Are Using Facebook to Organize for Civil War Amid Coronavirus. https://www.techtransparencyproject.org/articles/extremists-are-using-facebook-to-organize-for-civil-war-amid-coronavirus

[51] ISD (2020). Covid-19 Disinformation Briefing No.2. https://www.isdglobal.org/isd-publications/covid-19-disinformation-briefing-no-2/

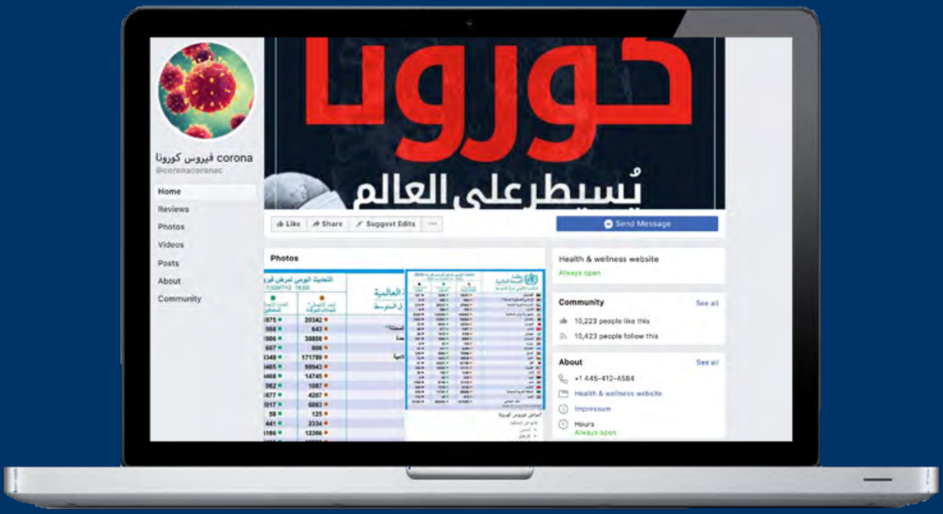


Figure 1: Arabic language 'Coronavirus' Facebook page from ISD research

identified itself under the 'health and wellness' category, boasting over 10,000 followers and managed by eight admins in Egypt, Syria and Yemen. It interspersed posts celebrating the US death toll from COVID-19 with pieces of overt ISIS content, including full-page uploads from the group's weekly newsletter al-Naba. In early April, the page's administrators issued a post stating that "everyone should be of the belief that this virus is a 'soldier of God' who is supporting His servants on Earth - the monotheists - whose chests are healed by the extensive death toll of infidels."[52]

200,000

posts across Twitter, Reddit, Tumblr, 4Chan and Voat contained the word 'boogaloo' between Feb and March 2020

155

pieces of ISIS video content shared using COVID-19 hashtags

570,000

number of views generated by ISIS video content using COVID-19 hashtags

WHAT IS THE IMPLICATION?

What is the implication for response? Following pressure from these and parallel studies, the Violence and Incitement policies on both Facebook and Instagram were updated on 1 May 2020, prohibiting 'the use of boogaloo terms when they are accompanied by statements and images depicting armed violence'. However, investigations conducted since the announcement[53] have shown the inadequacy of changes based solely on terminology. Extremist and hate groups are highly adaptive and find consistent workarounds for such policies, including code words, misspellings or symbols, which help them evade detection and continue spreading content online. These examples emphasise that moderation policies and enforcement must be iterative and proactive, learning from expert research on the evolving messages, symbols, tactics and vocabulary of known hate or extremist groups. They also demonstrate the need for enforcement to remain vigilant across different kinds of harm, despite the current focus on health-related disinformation that companies have, understandably, pivoted towards during the crisis. The consistent and comprehensive enforcement of Terms of Service across issue sets is vital to prevent gaps in response, including tangential threats that might not be immediately predictable or apparent (See: Recommendation Four).

[52] Full report on the 'Fuouaris Network' due to be published by ISD in late June.

[53] Egkolfopoulou, M. & Sebenius, A. (2020, May 12). Facebook Violence Curbs Thwarted by Groups Using Code Words. Bloomberg. https://www.bloomberg.com/news/articles/2020-05-12/facebook-violence-curbs-thwarted-by-groups-using-code-words

2. Advertising

Where was the policy position previously?

Over the past four years, platforms have answered calls for greater transparency on political advertising, most notably with the creation of publicly searchable archives and libraries. Political advertising has, for the most part, been narrowly defined as candidate-, party- or government-related content, except on certain platforms and in selected countries where 'political issue-based advertising' is included. These efforts have enabled greater visibility for the public and researchers on paid ad content and, to a limited extent, the associated targeting criteria, spend and buyers. Work from teams such as NYU CyberSecurity Center[54] and Mozilla Foundation[55] has monitored the successes and limitations of these efforts; last year ISD also reviewed platform commitments around political and issue-based ad transparency in Europe, exposing gaps in data provision for researchers and the public on both fronts.[56] Notably, these oversights led to abuses of paid advertising during the 2019 European Parliamentary elections. One of the critical gaps around advertising and disinformation is the exception Facebook makes for politicians, allowing them to use paid ads to promote falsehoods and preventing independent experts from fact-checking or labelling such content.

Despite these limits, companies have made significant progress in enabling more independent oversight of paid advertising on their platforms. There remain limits in both the narrowness of policies and the consistency of enforcement; nonetheless, the steps forward have greatly increased the capability of those seeking to expose disinformation and misrepresentation through paid content.

What has changed because of Coronavirus?

Facebook, Twitter and Google have each amended their ad policies to curb profiteering around COVID-19. This includes limits on the use of inflated prices and monopolies for Personal Protective Equipment (PPE) such as face masks, gloves and hand sanitiser, alongside the range of alleged cures, remedies, diagnostic tests and vaccines exploding onto the market since January 2020. Another raft of measures aims to support the wider public health and stability effort, with companies prohibiting ads that incite panic, contravene government advice or scapegoat certain individuals/groups. Platforms are also providing in-kind ad credit and promotion for expert bodies, including the World Health Organisation and national governments, helping them communicate updates quickly to the widest possible audience. YouTube has vacillated on policies that determine who can advertise about COVID-19; initial limits were strict, but the platform now allows a fairly broad range of actors to promote content concerning the pandemic.

For a full overview of platform responses, please refer to **Annex 1**.

[54] NYU CyberSecurity Center (2019). Online Political Ads Transparency Project. https://cyber.nyu.edu/2019/06/13/online-political-ads-transparency-project/

[55] Lloyd, P J. & Geurkink, B. (2019, July 19). Under the Hood: Mozilla's Fight for More Transparent Ads. Mozilla Foundation. https://foundation.mozilla.org/es/blog/under-the-hood-mozillas-fight-for-more-transparent-ads/

[56] ISD's full report - "Cracking the Code: An Evaluation of the EU Code of Practice on Disinformation" - is due for publication later this year.

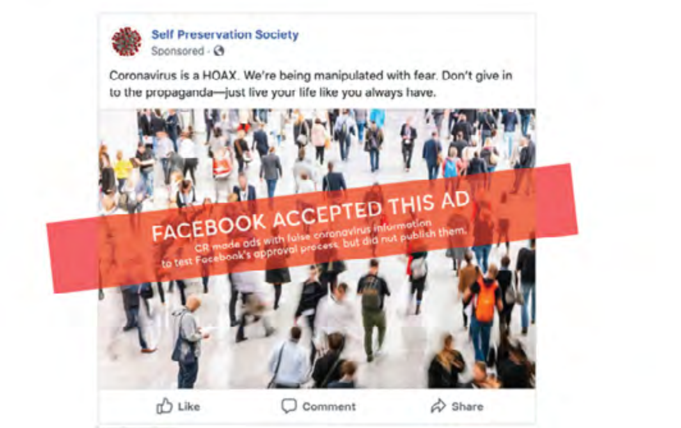
Advertising
Case Study 1

Research shows Coronavirus misinformation is not sufficiently vetted on Facebook ads[57]

Consumer Reports

Facebook has introduced systems to enforce advertising rules around COVID-19, but too many violations are slipping through the net. Measures are not sufficiently robust to prevent ads promoting hateful or false information about the pandemic, but it is unclear whether such errors relate to machine- or human-based decision-making.

In an attempt to test new policies[58] on coronavirus misinformation, researchers trialled the vetting on Facebook's paid advertising system. The organisation created a fake Facebook account and page to post seven ads on the platform, all directly violating the company's policies on false and dangerous information; the content 'remained scheduled for publication for more than a week without being flagged by Facebook', but Consumer Reports withdrew all ads from the queue to ensure they were never publicly viewed. In the wake of the test, Facebook confirmed that all seven ads had violated its policies, and included posts such as the below:



Self Preservation Society
Sponsored ·

Coronavirus is a HOAX. We're being manipulated with fear. Don't give in to the propaganda—just live your life like you always have.

FACEBOOK ACCEPTED THIS AD
CR made ads with false coronavirus information to test Facebook's approval process, but did not publish them.

Figure 2: One of the ads accepted by Facebook during Consumer Reports' study. Source: Consumer Reports.

WHAT IS THE IMPLICATION?

To prevent scams and other widespread harms caused by advertising, including disinformation, experts need maximum insight on the nature and scale of gaps. This includes rigorous transparency over advertising data (See Recommendation Two), as well as the processes used to allow or prohibit such content on the platform. The latter will require systemic transparency around companies' decision-making protocols and software, including those operated by humans versus machines (See Recommendation One).

[57] Waddell, K. (2020, April 7). Facebook Approved Ads With Coronavirus Misinformation. Consumer Reports. https://www.consumerreports.org/social-media/facebook-approved-ads-with-coronavirus-misinformation/

[58] Jin, K-X. (2020, June 11). Keeping People Safe and Informed About the Coronavirus. Facebook. https://about.fb.com/news/2020/06/coronavirus/#exploitative-tactics

Advertising **Case Study 2**

Chinese state-funded TV ads run without political disclaimers on Facebook

The Telegraph[59] and VICE[60]

The processes to make paid political activity on Facebook transparent to users are not properly or comprehensively enforced across regions and languages. This allows foreign states to pay for visibility and targeting without any disclosure on who or what entity is funding the service.

While stricter rules for political ad transparency have been implemented by Facebook in the US and some European countries, as described above, planned roll-outs in other regions have been delayed. Two reports by The Telegraph show how this has allowed Chinese state-funded media outlets to run targeted adverts on Facebook without disclaimers[61] about the financial source, including in some countries[62] where the new transparency measures are supposedly in place. An investigation by Vice[63] showed that Facebook and Instagram had both allowed undisclosed political ads to run from the Global Times, Xinhua News Agency, Global China Television Network (GCTN), and China Central Television (CCTV) about coronavirus, 'targeting users around the world in English, Chinese, and Arabic'.

WHAT IS THE IMPLICATION?

What is the implication for response? In June 2020, Facebook announced changes[64] to their policies that mean media outlets that are 'wholly or partially under the editorial control of their government' will be labelled as such across the platform, which will include ads from these publishers 'later this year'. Systemic transparency for advertising is a critical first step in preserving the integrity of financial spending around elections, and must be part of transparency regimes that enable oversight of how the systems and decisions of companies are affecting users and their rights (See Recommendation One).

[59] Dodds, L. (2020, April 5). China floods Facebook with undeclared coronavirus propaganda ads blaming Trump. The Telegraph. https://www.telegraph.co.uk/technology/2020/04/05/china-floods-facebook-instagram-undeclared-coronavirus-propaganda/

[60] Gilbert, D. (2020, April 7). China's Been Flooding Facebook With Shady Ads Blaming Trump for the Coronavirus Crisis. Vice. https://www.vice.com/en_uk/article/9397v8/chinas-been-flooding-facebook-with-shady-ads-blaming-trump-for-the-coronavirus-crisis

[61] Dodds, L. (2020, April 26). China exploits Facebook delays over advertising rules to spread coronavirus propaganda. The Telegraph. https://www.telegraph.co.uk/technology/2020/04/26/china-exploits-facebook-delays-advertising-rules-spread-coronavirus/

[62] Dodds, L. (2020, April 5). China floods Facebook with undeclared coronavirus propaganda ads blaming Trump. The Telegraph. https://www.telegraph.co.uk/technology/2020/04/05/china-floods-facebook-instagram-undeclared-coronavirus-propaganda/

[63] Gilbert, D. (2020, April 7). China's Been Flooding Facebook With Shady Ads Blaming Trump for the Coronavirus Crisis. Vice. https://www.vice.com/en_uk/article/9397v8/chinas-been-flooding-facebook-with-shady-ads-blaming-trump-for-the-coronavirus-crisis

[64] Gleicher, N. (2020, June 4). Labelling State-Controlled Media on Facebook. Facebook. https://about.fb.com/news/2020/06/labeling-state-controlled-media/

Advertising **Case Study 3**

Coronavirus scams enabled by Google and Facebook ads

Tech Transparency Project[65]

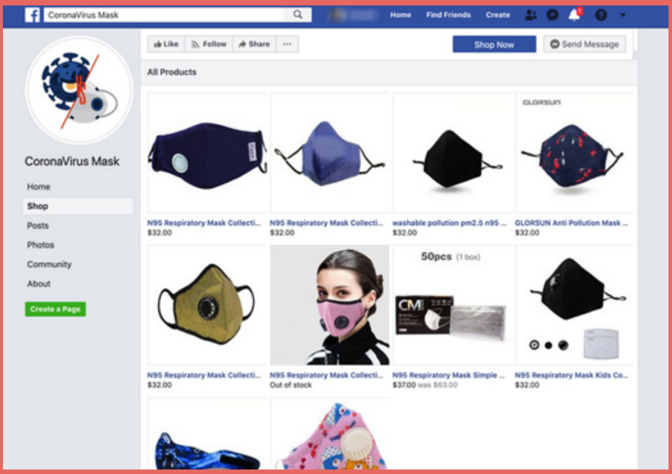

CoronaVirus Mask
Home
Find Friends
Create
Like
Follow
Share
Shop Now
Send Message
CoronaVirus Mask
Home
Shop
Posts
Photos
Community
About
Create a Page
All Products
N95 Respiratory Mask Collecti...
$32.00
N95 Respiratory Mask Collecti...
$32.00
washable pollution pm2.5 n95 ...
$32.00
GLORSUN Anti Pollution Mask ...
$32.00
N95 Respiratory Mask Collecti...
$32.00
N95 Respiratory Mask Collecti...
Out of stock
N95 Respiratory Mask Simple ...
$37.00 was $63.00
N95 Respiratory Mask Kids Co...
$32.00
50pcs (1 box)

Figure 3: Advertisements used to sell face masks on Facebook marketplace, identified in Tech Transparency Project's investigation. Source: Tech Transparency Project.

130+

Facebook pages selling medical face masks, including some labelled as N95 respirators

Despite companies' best efforts, actors are still able to profiteer off crises online, whether through ads or other commercial services such as marketplaces. This exposes users to potential scams and exploitation.

Facebook and Google have both failed to clamp down on the sale of and profiteering from medical face masks on their platforms, despite efforts to ban the practice. Research from Tech Transparency Project identified more than 130 Facebook pages selling medical face masks, including some labelled as N95 respirators. The pages were found through Facebook searches for terms like "corona mask," "N95," and "surgical mask." The report found similar examples on Facebook-owned platform Instagram, as well as face mask ads 'served by Google on a variety of websites'. Facebook itself is currently taking legal action[66] against a software company that it claims ran deceptive advertisements on the platform, including pointing users to 'investment scams and bogus information about the coronavirus pandemic'.

WHAT IS THE IMPLICATION?

The scale and impact of these gaps in enforcement can only be fully understood with broader, more granular data access, especially for researchers attempting to protect the public and expose incidents of fraud or harm (**See Recommendation Two**). Significant steps have been taken to improve the transparency of political and issue-based advertising across the three platforms; nonetheless, there remain ways to improve the quality and breadth of data access for independent oversight.

[65] Tech Transparency Project (2020, March 23). Tech Giants Fail to Eradicate Face Mask Sales. https://www.techtransparencyproject.org/articles/broken-promises-tech-giants-fail-to-eradicate-face-mask-sales

[66] Bloomberg News. (2020, April 10). FACEBOOK SUES OVER DECEPTIVE ADS LINKED TO CORONAVIRUS SCAMS. Ad Age. https://adage.com/article/digital/facebook-sues-over-deceptive-ads-linked-coronavirus-scams/2249641

3. Proactive information

Where was the policy position previously?

Social media platforms have, on occasion, proactively offered authoritative information during high profile events. To date this has included reminders to vote on election day, targeting users in relevant geographies on Facebook, and 'election labels' on Twitter that provide vetted partner information about US candidates, including the political office they are running for and the associated state.[67] Facebook also launched a Crisis Response Hub in 2017, which provides various tools to help users navigate online news and information during natural disasters, terrorist attacks or life-threatening incidents. Among these is a function for users involved or potentially affected to 'Get information', with proactive, curated updates about emerging and ongoing natural disasters.[68] However, the scale, longevity and global nature of the COVID-19 crisis has required an unprecedented response from platforms in providing authoritative information directly to users.

What has changed because of Coronavirus?

COVID-19 has sparked a major shift in tech companies' partnership with public sector and multilateral bodies worldwide. Hoping to address the explosion of misinformation around the virus - ranging from fake cures to false infection data - platforms are helping expert bodies to package and disseminate updates to the widest possible audience. This includes:

- Pop-up messages from the World Health Organisation and other key agencies, triggered by searches for 'coronavirus' or related keywords;
- Alert messages to users who have liked, reacted to or commented on content previously identified as false (e.g. by third-party fact-checkers);
- Public service announcements from health agencies on Newsfeeds, Homepages and Timelines;
- Newly launched portals and information hubs on COVID-19, collating news from credible sources, data insight, travel guidance, links to relief efforts, official prevention tips and health guidance;
- Platform campaigns to promote health advice, including Stay At Home orders and personal hygiene;
- Direct messaging systems for users to interact with health agencies, including WHO (both automated and personalised);
- Education resources to promote Media and Digital Literacy;
- Guides to stay safe online, including for misinformation and health-related scams.

Platforms are also trying to maximise their messaging apps for public outreach and education. Facebook Messenger has introduced a function that allows users to directly contact the WHO for information, providing quick answers to queries on the pandemic. The Messenger Coronavirus (COVID-19) Community Hub[69] is also active and designed to help people educate their friends, family and peers, with some guidance on how to avoid scams and pandemic mis/disinfo. In parallel, WhatsApp launched their 24-hour WHO Health Alert[70], a tool updated daily which responds to key words around prevention, travel advice and common myths, and a Coronavirus Information Hub[71] in partnership with the WHO, Unicef, UNDP and the Poynter International Fact-Checking Network.

A full overview of platform responses can be found in **Annex 1**.

[67] Why am I seeing a reminder about an election and voting on Facebook?. Facebook.
https://www.facebook.com/help/1519550028302405
Coyne, B. (2018, May 23). Introducing US Election Labels for Midterm Candidates. Twitter.
https://blog.twitter.com/en_us/topics/company/2018/introducing-us-election-labels-for-midterm-candidates.html

[68] Crisis Response Hub. Facebook. https://www.facebook.com/about/crisisresponse/

[69] https://www.messenger.com/coronavirus

[70] Whatsapp. The World Health Organization launches WHO Health Alert on WhatsApp. Whatsapp. https://www.whatsapp.com/coronavirus/who

[71] https://www.whatsapp.com/coronavirus

What is the evidence around impact?

Companies have been quick to implement tools that promote credible and trustworthy information to users in public spaces on their platforms. While efforts to amplify facts are laudable, there is little data on the impact of banner ads, pop-up messages or information boxes on user attitudes, and therefore little attempt to assess their merits and pitfalls. Moreover, private groups on Facebook do not contain any such measures for proactive information, despite repeated[72] evidence that they are awash with COVID-19 disinformation.

To gauge the relevant impact and limitations of such measures, we need defined metrics developed by experts in online behaviour and made public by platforms. Facebook have announced initial click-through rates on some content (referenced in Annex 1), but this information alone can be deceptive - it does not indicate how long users spent reading verified information or, above all, whether it affected their overall stance on the crisis. Indeed, research from ISD has shown that engagement rates with known disinformation on Facebook dwarfs that of credible sources (see Case Study One). Pilot studies are critical to answering key questions about whether those viewing proactive content are less likely to engage with or trust mis- and disinformation online.

[72] Scott, M. (2020, March 30). Facebook's private groups are abuzz with coronavirus fake news. Politico. https://www.politico.eu/article/facebook-misinformation-fake-news-coronavirus-covid19/

Recommendations: Building Systems Resilient to Disinformation

While the coronavirus pandemic may be rare in terms of its severity and reach, the information crisis that has emerged in parallel speaks to wider issues: namely, a digital system underprepared to deal with viral lies and manipulation of identity, attention, and popularity.

There is a tendency to focus on content removal when discussing companies' efforts around disinformation; however, the evidence above confirms that addressing actors, behaviours and distribution mechanisms is equally vital and often overlooked. Our recommendations broach various elements in the infrastructure of disinformation, suggesting immediate steps to help mitigate threats and limit both the capacity and impact of those intending to cause harm.

ISD has developed **six recommendations** for social media companies, intended to better protect users against disinformation writ large, whether COVID-related or otherwise. They suggest immediate steps for companies to improve the resilience of their platforms against manipulation, particularly in crisis moments, and to enable more consistent, transparent and robust enforcement of existing policies. Such efforts could be adopted independently of emerging regulation in the EU, UK and elsewhere, but sit within ISD's broader call for democratic governments to take sustained and rigorous action, improving platform accountability on threats to public safety enabled or created by their products. In the context of liberal democracies, ISD has supported the nascent design of regulation measures[73] in order to enforce transparency for advertising, content moderation, enforcement protocols, appeals and redress practices, as well as for processes of algorithmic decision-making and their outcomes.[74]

Democratic oversight will, in the end, be required to assess the efficacy of any voluntary approaches, which at present are largely reactive and limited in scope. The nature of that oversight may vary from context to context, but the core principle must remain the same: transparency. Platforms that have become public spheres must make those spaces as intelligible as possible, both for users, experts and those elected to maintain public safety. Social media play an increasing role in shaping our culture, informing our political decision-making, and driving societal change; as such, the activities that define their usage should be open and observable, both computationally and otherwise. Transparency is the precondition for establishing trust in information systems, during crisis moments and beyond: there remains too much that is invisible to users, including how and why they experience information as they do.

[73] ISD. (July 2019). Extracts From ISD's Submitted Response to the UK Government Online Harms White Paper. https://www.isdglobal.org/isd-publications/extracts-from-isds-submitted-response-to-the-uk-government-online-harms-white-paper/

[74] ISD. (April 2020). Algorithm Inspection and Regulatory Access. https://www.isdglobal.org/isd-publications/algorithm-inspection-and-regulatory-access/

1. Moving beyond content-driven approaches: Increase transparency and user control over distribution mechanisms.

The crisis of disinformation around COVID-19 is not an issue of false information alone. The often binary discussion over content removal versus freedom of speech obscures the fundamental role that distribution mechanisms play in amplifying and targeting content beyond its original audience. Efforts to counter harmful disinformation should focus as much on the channels of distribution as the nature of content itself.

These mechanisms, be it the micro-targeting of ads or using algorithms to recommend the next piece of visible content, constantly make decisions for users about what they can see online. They also play an intrinsic role in the disinformation ecosystem, amplifying[75] dangerous content that might otherwise reach a limited audience.

Platforms' business models dictate the success or failure of certain content, and have been shown[76] to favour sensationalist, divisive or controversial posts in order to increase traffic and user engagement, and therefore maximise advertising revenue. Labelling debunked content has helped alert many users to false information spreading on their channels and feeds, but reactive labelling alone cannot counter systems that incentivise clickbait and emotion over authority or evidence. There is enormous work required to deconstruct companies' emphasis on popularity over accuracy when curating information, and to grant users more control over how their information is selected, ordered, and recommended.

To be truly effective, democratic governments will need to mandate transparency from platforms through regulation. In the interim, there are actions that companies themselves can take to better protect users from harm. To begin rectifying the power of opaque systems, companies should provide users with greater autonomy over what populates their social media feeds and why; this includes options based on the chronological ordering of content, or recommendation systems that prioritise trusted, authoritative sources when selecting and promoting information. 'Stage-gates' on potential disinformation could also be established to reduce the risk of viral 'waves', as outlined by Google News creator Krishna Bharat[77]. This might include the insertion of automatic breaks in rapid news spikes, allowing verification and human vetting before a story gains traction. Such measures would help limit intentional disinformation campaigns, verifying stories or claims before they spread beyond a tipping point of reach and engagement online.

[75] Diresta, R. (2018, August 30). Free Speech Is Not the Same As Free Reach. Wired. https://www.wired.com/story/free-speech-is-not-the-same-as-free-reach/

[76] Horwitz, J. & Seetharaman, D. (2020, May 26). Facebook Executives Shut Down Efforts to Make the Site Less Divisive. Wall Street Journal. https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499

[77] Bharat, K. (2017, April 27). How to Detect Fake News in Real-Time. Medium. https://medium.com/newco/how-to-detect-fake-news-in-real-time-9fdae0197bfd

2. Unpacking online content and behaviours: Support good-faith research through privacy-protected data and insights on detection methods, both for inauthentic and coordinated behaviour.

Independent researchers and academics are continually working to expose platform manipulation on social media, both in crisis situations and beyond, but face growing restrictions on data access. Moreover, the fast-evolving tactics of actors engaging in such activity are often hard to model, especially for research bodies that lack any reasonable data on the activities of these perpetrators. Such data is crucial even after companies have removed the offending content from their platforms, as it can aid longitudinal trend analysis and open-source investigations. Independent experts need this to assess if and how platforms are being weaponised, whether to deceive users or distort the available flow of information.

Where they do not already exist, for example through Twitter's API, companies must develop effective protocols for privacy-protected data, improving access for vetted and independent research bodies. The Social Science One[78] experiment offers some precedent: the scheme attempted to share large datasets relevant to the study of disinformation, safely and without compromising either user privacy or business competition. Lessons must be learned from that effort to avoid similar obstacles in the future, for example:

- The legal and technical complexity of data-sharing efforts should not be underestimated.
- The effort has catalysed a new framework for 'scholarly and ethical review of networked data research', providing industry standards that should be considered in future data-sharing models.[79]
- The statistical method they developed for differential privacy should be referenced to design data-sharing systems that can 'preserve the privacy of end users while enabling scholars to draw valid statistical inferences on the questions they are investigating'.[80]

In the near term, there are creative routes for increasing researchers' capability, helping them detect disinformation campaigns without formal regulation or government oversight. Companies have a chance to take the initiative, using their privileged insight into which signals help detect co-ordinated disinformation on their platforms. New collaborations should be trialled in this vein, whereby they produce 'dummy data' that artificially simulates cases of platform manipulation. By creating invented scenarios you mitigate risks to data privacy and competition, but still improve knowledge-sharing with the independent research sector and enable large-scale quantitative data analysis.

[78] https://socialscience.one/

[79] Social Science One (2020, February 1). Analyzing Data From Facebook. https://socialscience.one/blog/analyzing-data-facebook

[80] Evans, G. and King, G. (2020, May 16). Statistically Valid Inferences from Differentially Private Data Releases, with Application to the Facebook URLs Dataset. Harvard. https://gking.harvard.edu/dpd

3. Understanding the perpetrators: Pay as much attention to the full spectrum of non-state disinformation actors, as to those associated with foreign governments.

To address the current threat, platforms need to widen their lens around which groups are weaponising disinformation online. Extensive research from ISD[81] and other organisations has shown the range of actors seeking to deceive and divide audiences during the pandemic, corroborated by the companies' own detection teams[82]. This emerging evidence indicates that coordinated disinformation efforts go far beyond foreign state activity. What is seen as the Kremlin's Internet Research Agency (IRA) toolkit is already in the hands of hate groups[83], political parties[84] and conspiracy theorists[85], each of which pose different kinds of challenges to civil rights, public health and democratic processes.

A narrow focus on foreign states as perpetrators, still communicated in public statements[86] from some platforms, reveals a limited appetite to confront the numerous sources of disinformation, whether individuals, groups or coordinated networks within and across borders. Despite positive steps in recent months, more concerted efforts are needed to understand and address the range of actors at play. Companies will require a breadth of internal and external expertise to tackle this issue, covering the varied geographic and ideological identities of those involved in deceptive practices online. Companies should encourage and support good faith research to analyse non-state disinformation networks, a model which has some proven success: recently in Georgia[87], local fact-checking and reporting groups were able to provide insights directly to Facebook, prompting the detection and removal of coordinated inauthentic behaviour from a domestic political party and a private media firm.

[81] ISD. (2019). Interim Briefing: Propaganda and Digital Campaigning in the EU Elections. https://www.isdglobal.org/isd-publications/interim-briefing-propaganda-and-digital-campaigning-in-the-eu-elections/

[82] Facebook. (2020, May 5). April 2020 Coordinated Inauthentic Behavior Report. Facebook. https://about.fb.com/news/2020/05/april-cib-report/

[83] ISD. (2019). Interim Briefing: Propaganda and Digital Campaigning in the EU Elections. https://www.isdglobal.org/isd-publications/interim-briefing-propaganda-and-digital-campaigning-in-the-eu-elections/

[84] Facebook. (2020, May 5). April 2020 Coordinated Inauthentic Behavior Report. Facebook. https://about.fb.com/news/2020/05/april-cib-report/

[85] Graphika. (2020, May 5). Facebook Downs Inauthentic Cluster Inspired by QAnon. Graphika. https://graphika.com/reports/facebook-downs-inauthentic-cluster-inspired-by-qanon/

[86] Jack, S. (2020, May 21). Facebook's Zuckerberg defends actions on virus misinformation. BBC News. https://www.bbc.co.uk/news/business-52750162

[87] Facebook. (2020, May 5). April 2020 Coordinated Inauthentic Behavior Report. Facebook. https://about.fb.com/news/2020/05/april-cib-report/

4. Streamlining application: Create and enforce disinformation policies consistently across issues and actors to prevent gaps in response.

COVID-19 should not be addressed in a policy vacuum - while the crisis has posed acute challenges to platforms, the mechanisms to bolster prevention, mitigation and removal of content could apply to many disinformation threats. Indeed, as the pandemic evolves there is mounting evidence to link health and political disinformation, blending into a larger ecosystem of online harms. To address each emerging issue or crisis in isolation is impractical, and could hinder greater economies of scale in both technology and human resources. Platform manipulation policies, and their enforcement, should be transparent for users and consistent across different topics, in order to address disinformation in a sustained manner. This is particularly true given the exploitation of COVID-19 by malign actors, whether through doxxing, direct threats to individuals, hate speech, scams, or even co-opting trending topics to spread extremist material (as detailed above). A failure to tackle existing conspiracy and disinformation networks such as the QAnon and anti-vax movements has provided fertile ground for COVID-19, since these groups, channels and pages have a ready-made audience vulnerable to health disinformation. Siloed responses risk missing the bigger picture of weaponised information and fail to address hybrid threats.

Consistency should not, however, come at the expense of issue-specific expertise within companies' moderation and policy teams. In fact, the ability to design and enforce measures across areas of potential harm will rely on whether they can build in-house and external expertise into all decision-making. Understanding these threats is a prerequisite to enacting sensible and consistent content moderation online.

5. Addressing the ecosystem: Formalise cross-platform initiatives to address disinformation crises, similar to the models in place for terrorist incidents or child sexual exploitation online.

Facebook, Google, LinkedIn, Microsoft, Reddit, Twitter, and YouTube issued a joint statement around COVID-19 in March, stating the companies were 'working closely' on response efforts and particularly the fight against fraud and misinformation[88]. We have also seen specific collaborations on wider tech responses to the pandemic, including between Google and Apple around Bluetooth contact tracing[89]. However, beyond ascribing to a common message, it is unclear what these joint efforts involve or whether there has been formal, cross-platform action on disinformation.

Lessons should be taken from two existing initiatives. Firstly, the Global Internet Forum to Counter Terrorism (GIFCT)[90], a body created by Facebook, Microsoft, Twitter and YouTube in 2017 to disrupt terrorist abuse of their platforms and provide a mechanism for industry collaboration with experts, civil society and governments. A further five companies have joined since its foundation, and GIFCT also provides support for smaller, non-member companies. The forum has driven development of shared technology, information pathways and research to mitigate these risks, including a shared database of 'hashes' (digital fingerprints) for known terrorist content. Since the terrorist attacks in Christchurch in March 2019, GIFCT has also trialled common crisis response systems, especially for instances where terrorist-related content is shared during or in the wake of an attack. The model is far from perfect - their database still lacks public oversight, and it remains unclear how content decisions are made, including on government-ordered takedowns. Nonetheless, the forum has at least tabled an agenda and begun to articulate some common objectives.

Secondly, PhotoDNA[91], a tool created by Microsoft and made freely available to developers, helping detect and report material flagged as child sexual abuse. PhotoDNA has also created hashes which can be used across platforms, facilitated by key groups like the Internet Watch Foundation[92] in the UK, and WeProtect[93] globally.

Policy divergence between companies can prevent such fora from achieving their full potential, but this should not deter future efforts - the cross-platform nature of disinformation threats has long been evidenced and calls for more formal partnership. This could include crisis response protocols[94], and a hash-sharing database for known debunked content or manipulated video/images. Moreover, cross-sector support may enhance the ability of emerging, smaller and less well-resourced companies to confront disinformation on their platforms, embedding a response beyond the tech giants who often dominate discussion and research.

[88] Shu, C. & Shieber, J. (2020, March 17). Facebook, Reddit, Google, LinkedIn, Microsoft, Twitter and YouTube issue joint statement on misinformation. Techcrunch. https://techcrunch.com/2020/03/16/facebook-reddit-google-linkedin-microsoft-twitter-and-youtube-issue-joint-statement-on-misinformation/?guccounter=1

[89] Apple. (2020, April 10). Apple and Google partner on COVID-19 contact tracing technology. Apple. https://www.apple.com/uk/newsroom/2020/04/apple-and-google-partner-on-covid-19-contact-tracing-technology/

[90] https://gifct.org/about/

[91] https://www.microsoft.com/en-us/photodna

[92] https://www.iwf.org.uk/

[93] https://www.weprotect.org/

[94] https://www.gifct.org/joint-tech-innovation/

6.

Cutting through the noise: Leverage influencer networks and communication methods to share, amplify and surface verified content.

As ISD's research shows, verified and credible information is available on these platforms, but often fails to achieve the same level of reach or engagement as viral mis-/disinformation. In the case of COVID-19, this imbalance has proven true despite efforts by companies to promote content from organisations like the WHO. Social media is increasingly saturated, with millions of voices and ideas vying for attention, and as such there is less of a premium on fact than on format and messenger. The content which gains traction has increasingly high 'brand' values, both in terms of the author's status and the aesthetic medium itself - this is particularly true for emerging and more youth-focussed platforms like TikTok, where videos are designed for maximum virality, trend-setting and imitation.

To date, health experts have not seen real cause to enter this world of communication, and therefore lack creative tools to liaise with the public[95]. They now face an uphill battle to be heard and adapt to the new reality of a crowded, constantly-updating information landscape. To command their share of attention, institutions must be equipped with the tools and techniques of digital influencers, including everything from memes and bold visuals to slogans and gamified content. These are well-worn tactics for disinformation actors and conspiracy theorists, who understand the premium of grabbing and holding a user's gaze. Companies should provide guidance to health authorities and in-house support from brand and marketing teams, helping them package messages to capture the widest possible audience.

In addition, greater effort should be made to engage influencers themselves. The most popular YouTube channels have upwards of 140m subscribers, and every platform boasts its own roster of high-visibility, high-reach accounts. Moreover, these influencers often become 'credible messengers' for a particular sub-group or demographic, whether based on age, geography, ethnicity, gender, political affiliation or other unifying factors. This dynamic could provide novel points of entry to diverse audiences, many of whom are harder to reach via traditional methods or outlets. For such a model to work, tech platforms must provide the incentive structures for their networks to engage, and help broker direct dialogue with official agencies and institutions. This could include:

- Special briefing sessions for influencers by expert bodies, raising awareness on key mis/ disinformation trends and the associated dangers (especially in times of crisis);
- Twinning programmes between influencers and experts (e.g. public health officials, frontline service providers, victims of hate, former conspiracy theorists), providing a 'human face' to events and encouraging a more direct and personal transfer of information;
- Digital Literacy workshops for influencers, highlighting the tools and techniques to assess whether online content is credible and keep themselves/their followers well-informed (e.g. reverse image searches, signals of inauthentic behaviour, common traits of mis-/ disinformation, trusted fact-checking platforms, legal definitions of free speech and hate speech, available reporting mechanisms, advice and support helplines, relevant plug-ins and software, useful learning resources);
- Shared campaigns to promote key, verified information, prioritised by platform algorithms to increase exposure;
- Co-designed counter-messaging and broader digital citizenship content, in partnership with subject experts and third-party providers.

[95] DiResta, Renée (2020, May 6). Virus Experts Aren't Getting the Message Out. The Atlantic https://www.theatlantic.com/ideas/archive/2020/05/health-experts-dont-understand-how-information-moves/611218/

Conclusion: What Next?

Coronavirus has been a sobering moment in the fight against disinformation, forcing tech companies to reassess whether their policies and enforcement are fit-for-purpose. The response has varied considerably and been hampered at least in part by the logistical challenges of the pandemic, including furloughed staff and remote working.

The cumulative evidence in this paper suggests that policies around advertising have been more rigorously enforced during the crisis, when compared with those surrounding unpaid user generated content, algorithmic distribution of disinformation domains and outlets, or the scaled fact-checking and labelling of misleading, false and harmful content. Such findings may have multiple explanations. The relatively small scale of advertising content makes detection and removal an easier task than with unpaid user posts or activity. Consistent pressure to improve transparency on political ads has also been a factor, allowing more robust audits of companies from the research community and increased scrutiny of failures over the past four years. This has helped identify areas for adapted policies and better enforcement, and may also act as a deterrent for those seeking to exploit platform services for harm. **Transparency therefore straddles both prevention and response - dis-incentivising efforts from the outset, and enabling the identification and troubleshooting of issues as they arise.**

Sadly, any conclusions drawn must rely on some element of extrapolation and inference. Without better access to data and insight on companies' decision-making systems, both human- and machine-led, we cannot determine with certainty why some areas of policy appear more effective or better enforced than others. The disinformation incidents outlined above were exposed despite minimal data access - one can only imagine the real scale of the problem on those platforms, or what could be achieved with more candid partnerships between the tech and research sectors.

At some point, the COVID-19 crisis will end or become a managed part of public health systems worldwide. It would be naive to assume the so-called 'infodemic' will follow suit, or that company systems will become more resilient on their own. We must learn from the acute challenges of this moment and the flaws it has exposed in our ability to prevent, identify and counter disinformation online.

ANNEX 1:
Platform Responses to COVID-19

Accurate as of 8th June 2020

Twitter[96]

In March, Twitter broadened its definition of harm to 'address content that goes directly against guidance from authoritative sources of global and local public health information'. This includes denying established facts about COVID-19, having a fake call to action that benefits a third party, creating panic based on fake claims, impersonating government health officials, circulating false diagnostic advice, and promoting claims about the immunity or susceptibility of certain groups. In May, the company announced it would label Tweets containing COVID-19 misinformation with the following disclaimer: "some or all of the content shared in this Tweet conflicts with guidance from public health experts regarding COVID-19". This could include applying a Public Interest Notice to world leaders who violate the guidelines, although such action has been extremely rare to date. They note the difficulty in taking 'enforcement action' on every Tweet containing incomplete or disputed information about COVID-19, citing reduced capacity in the moderation team. This considered, a #KnowTheFacts search prompt aims to guide users to credible information, e.g. from WHO or national health agencies, and can detect common misspellings of keywords. At the time of writing, this initiative is active in 70 countries worldwide. The platform has also updated its Event feature to ensure credible updates on the pandemic, which appears at the top of the Home timeline for users in 30+ countries.

At a broader level, Twitter have supported or partnered with national NGOs as part of the International Fact-Checking Network (IFCN). These third-party bodies work in real-time using data journalism techniques, both to identify and report coordinated campaigns and debunk the mis/disinformation therein. In February 2020 the platform launched new labels[97] for Tweets containing synthetic or manipulated media - similar flags will now appear on Tweets with harmful or misleading content around COVID-19, applied retroactively to anything on the site. The labels link to a Twitter-curated page of external 'trusted sources', and may be supplemented with a warning that informs users the Tweet conflicts with public health advice. They categorise and respond to such content under three areas: misleading information (labelled or removed), disputed claims (labelled and warning) and unverified claims (no action, unless it contravenes the new Content Policy).

In regards to advertising, Twitter is offering Ads for Good credits to governments as well as targeting racketeers. They explicitly mention crackdowns on:

- Distasteful references to COVID-19;
- Content likely to incite panic;
- Inflated prices for health-related products (e.g. PPE, alcohol hand sanitizer).

References to vaccines, treatments and test kits are only permitted from news publishers exempted under the Political Ads Content Policy, and must be informational.

[96] Twitter. (2020, April 2).Our ads policy for COVID-19. Twitter. https://blog.twitter.com/en_us/topics/company/2020/covid-19.html#adspolicy

[97] Roth, Y. & Achuthan, A. (2020, February 4). Building rules in public: Our approach to synthetic & manipulated media. Twitter. https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html

Facebook[98]

Facebook has adapted its moderation policies, incorporating false claims and conspiracies flagged by health bodies under their existing Content Violation measures. The company is partnering with the WHO, providing 'as many free ads as they need' to circulate timely and accurate updates on the pandemic, alongside in-kind ad credits to public health organisations and global health experts. The WHO have also launched an interactive experience on Messenger, providing quick answers to individual queries on the pandemic. Combined with the COVID-19 Information Centre[99] and pop-ups on Facebook/Instagram, they claim to have directed 2 billion people to resources from health authorities, with 350 million users 'clicking to learn more'. The Info Centre includes a 'Get the Facts' tab with articles selected by the News curation team and updated weekly. They have banned all ads and commerce listings for face masks, hand sanitizer, surface disinfecting wipes and COVID-19 testing kits.

On the monitoring side, the platform has officially partnered with 60 fact-checking organisations operating in 50 languages worldwide. In April the platform announced it was 'sending content reviewers home', which has led to increased use of automation, amended prioritisation for user reports, and temporary changes to the appeals process. Some full-time employees are reviewing content related to 'real-world harm[100] (e.g. Child Safety, Suicide and Self-Harm, Terrorism), but a wholescale return to work is not foreseeable in the short term. According to their own website[101], Facebook 'expect to make more mistakes, and reviews will take longer than normal'. Nonetheless, the platform reports that it placed warning labels on 50 million pieces of content in April alone[102], based on 7,500 articles by its independent partners. This is intended to reduce the distribution of flagged content and prompt similarity detection methods, helping identify duplicates of debunked stories. New alert messages will appear in the Newsfeeds of people who have liked, reacted to or commented on known misinformation, connecting people to WHO myth-busting. COVID-19 related groups will also be shown these messages, and admins will be prompted to share Live broadcasts by bodies like the Centres for Disease Control and Prevention (CDC) and national health agencies. More than 2,000 state and municipal actors have been onboarded to Facebook Local Alerts, helping them communicate directly with their citizens, while the Messenger Coronavirus (COVID-19) Community Hub[103] is designed to help people educate their friends, family and peers, with some guidance on how to avoid scams or pandemic mis/disinfo.

Facebook have also announced the first 8 recipients of a $1m grant programme[104] with the International Fact-Checking Network, including grantees from France, Indonesia, Canada, Jordan, Kenya, Taiwan, Ukraine and Australia, and are due to launch further projects soon. This comes alongside an added $100m investment to support the news industry, including $25m of emergency grant funding for local journalists and $75m in marketing spend to outlets struggling to generate revenue.

[98] Jin, K., Keeping People Safe and Informed About the Coronavirus. Facebook. https://about.fb.com/news/2020/06/coronavirus/

[99] https://www.facebook.com/coronavirus_info/

[100] https://about.fb.com/wp-content/uploads/2020/03/March-18-2020-Press-Call-Transcript.pdf

[101] Jin, K., Keeping People Safe and Informed About the Coronavirus. Facebook. https://about.fb.com/news/2020/06/coronavirus/

[102] Rosen, G. (2020, April 16). An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19. Facebook. https://about.fb.com/news/2020/04/covid-19-misinfo-update/

[103] https://www.messenger.com/coronavirus

[104] Goldshlager, K. & Watson, O. (2020, April 30). Launching a $1M Grant Program to Support Fact-Checkers Amid COVID-19. Facebook. https://www.facebook.com/journalismproject/coronavirus-grants-fact-checking

WhatsApp

In March the platform launched a 24-hour WHO Health Alert[105], activated by typing 'hi' in a message to +41 79 893 1892. The service is updated daily with the latest information and responds to key words around prevention, travel advice and common myths. The system was launched in English with aims to expand into the 5 other official UN languages (Arabic, Chinese, French, Russian and Spanish). WhatsApp have also launched a Coronavirus Information Hub[106], in partnership with the WHO, Unicef, UNDP and the Poynter International Fact-Checking Network. The Poynter network has received a $1m grant to support the #CoronaVirusFactsAlliance, expanding the presence of local fact-checkers and monitors on WhatsApp.

Instagram[107]

Instagram has sought to improve public awareness around the pandemic, including education resources in Instagram Search, tags to promote accurate information and a "Stay Home" sticker for shared stories on social distancing. Other stickers introduced in Stories are meant to help share information on virus prevention (e.g. hand-washing), and the app will suggest content from credible experts and organisations if you search a related hashtag. Any COVID-19 accounts have been removed from their 'Recommendations' function, unless posted by a known health agency, and some additional content should be taken down from Explore. Content will be downranked in Feed and Stories if rated false by third-party fact-checkers[108], and removed entirely from Explore and hashtag pages. This includes false claims or conspiracy theories flagged by the science community. New measures also prohibit ads and commerce listings that refer to COVID-19 in order to 'create urgency, guarantee cures or prevent people from contracting it', alongside a temporary wholesale ban on content promoting medical supplies. There is a dedicated channel for local governments to share listings they believe violate local laws. A pop-up box now appears for searches around coronavirus, directing users to the WHO, Unicef and local ministries/international agencies.

Google

Have committed an initial $50m to the global COVID-19 response, although primarily to support humanitarian response, economic relief and recovery and distance learning. It is unclear whether funds have been earmarked to bolster efforts against mis/disinformation, although various new platforms and services have emerged since January 2020. This includes the COVID-19 Information and Resources[109] portal which houses data on the pandemic, safety and prevention tips, trending news, links to support relief efforts and other resources. Google have established a 24-hour incident response team to stay in sync with health authorities, and are optimising their search engine to provide accurate updates on the pandemic. This includes the SOS Alert[110] which siphons news from credible sources such as the WHO, government agencies, scientific journals and media outlets, with country-specific tabs on Travel, Closures, Testing, Reopening and General Information. They are also building a Knowledge Panel for COVID-19 (an information box which appears when users search for related symptoms, treatment or prevention methods), and are blocking thousands of ads which attempt to profit on the crisis via Google Ads (e.g. vendors selling masks at inflated prices, or offering fake cures and remedies). The company's Trust and Safety Team are helping to run PSAs on the crisis alongside health and government agencies.

[105] Whatsapp. The World Health Organization launches WHO Health Alert on WhatsApp. Whatsapp. https://www.whatsapp.com/coronavirus/who

[106] https://www.whatsapp.com/coronavirus

[107] Instagram. (2020, March 24). Keeping People Informed, Safe, and Supported on Instagram. Instagram. https://about.instagram.com/blog/announcements/coronavirus-keeping-people-safe-informed-and-supported-on-instagram

[108] Instagram. (2019, December 16). Combatting Misinformation on Instagram. Instagram. https://about.instagram.com/blog/announcements/combatting-misinformation-on-instagram/

[109] https://www.google.com/intl/en_uk/covid19/

[110] https://www.google.com/search?q=coronavirus

YouTube

On YouTube developers are trying to forefront authoritative sources, especially via the Top News shelf (highlighting videos from news outlets in search results) and the Breaking News shelf (highlighting videos about current events on the YouTube homepage). Content in the latter case is populated algorithmically, using signals including relevance to COVID-19, how up-to-date a story is, and the respective region. The platform have implemented new information panels[111] that drive users to the WHO or other resources in line with local guidelines, triggered by any COVID-related search term and available in 28 markets with Arabic, Spanish, French and English functionality. They are also partnering with health authorities and medical practitioners to release proactive content[112], donating ad inventory to accelerate messaging across the network. In April the platform introduced new features in the Explore tab to promote creators participating in the #StayHome and #WithMe campaigns (now running in 17 markets), and added a link to self-assessment tools in the COVID-19 search panel.

YouTube have also updated their Community Guidelines to include a COVID-19 Misinformation Policy[113], which outlines the removal of any videos denying the existence of the virus, promoting fake remedies or diagnostic tests, spreading conspiracies about its origin, claiming certain races/ethnicities are immune, discouraging people from seeking medical advice, linking the pandemic to 5G, undermining social distancing, or scapegoating groups (e.g. Asian communities) for the spread of disease. They have expanded monetization of COVID-related content to all creators and news organisations, assuming it follows Community Guidelines and is duly fact-checked. As the pandemic falls under their 'sensitive events' policy, YouTube will start enabling ads discussing COVID on a limited number of channels, including creators who accurately self-certify and various news partners. Only public sector ads are whitelisted against searches for 'coronavirus'. Due to staff reductions there will be increased use of automated content review (e.g. AI software, machine learning) over human-led moderation - this relies in large part on videos being flagged to the system via their standard reporting tool.

.

[111] https://support.google.com/youtube/answer/9004474

[112] Nguyen, N. [NHS]. (2020, March 6). Coronavirus - common questions | NHS. Youtube. https://www.youtube.com/watch?v=QV_UnPl8qMA

[113] https://support.google.com/youtube/answer/9891785

About the Institute for Strategic Dialogue

The Institute for Strategic Dialogue (ISD) is an independent nonprofit organisation dedicated to safeguarding human rights and reversing the rising tide of hate, extremism and polarisation worldwide.

We combine sector-leading expertise in global extremist movements with advanced digital analysis of disinformation and weaponised hate to deliver innovative, tailor-made policy and operational responses to these threats.

ISD draws on fifteen years of anthropological research, state-of-the-art digital analysis and a track record of trust and delivery in over 40 countries around the world to:

- Support central and local governments in designing and delivering evidence-based policies and programmes in response to hate, extremism, terrorism, polarisation and disinformation.
- Empower youth, practitioners and community influencers through innovative education, technology and communications programmes.
- Advise governments and tech companies on strategies to mitigate evolving online harms and achieve a 'Good Web' that reflects liberal, democratic values.

Only in collaboration with all of these groups can we hope to outcompete extremist mobilization and build safe, free and resilient societies for generations to come

All of ISD's programmes are delivered with the support of donations and grants. We have the data on what works. We now need your help to scale our efforts.

Support our work: isdglobal.org/donate

General
info@isdglobal.org

Media
media@isdglobal.org

Support
getinvolved@isdglobal.org

ISD | Powering solutions to extremism and polarisation

PO Box 75769 | London | SW1P 9ER | UK
www.isdglobal.org