FANNING THE FLAMES:

Foreign State-Sponsored Disinformation in the Time of COVID

DISINFO CLOUD

MAY 2021

Authored by

Rafia Bhulai, Christina Nemr, Marine Ragnet, and Eliza Thompson

Acknowledgements

The authors are grateful for the contributions and review of this report by Gabriel Delsol, Brendan O'Donoghue, and Andrew Rothgaber of Park Advisors, as well as the editorial assistance provided by Rhonda Shore.

This report was produced with support from the U.S. Department of State's Global Engagement Center. Any views expressed herein are those of the authors and do not necessarily reflect the views of the U.S. Department of State and Park Advisors. Any errors contained in this report are the authors' alone.

0. Table of Contents

1. Introduction

In March 2019, we released a publication titled "Weapons of Mass Distraction: Foreign State-Sponsored Disinformation in the Digital Age." That report provided an interdisciplinary review of the human and technological vulnerabilities to foreign propaganda and disinformation as well as an overview of several countries notable for such operations, including Russia and China.[1] The research surveyed in that report, particularly around human and technological vulnerabilities to disinformation, still holds true. Between March 2019 and March 2021, however, the world changed.

The emergence of COVID-19 in 2020 introduced not only a devastating virus but a vacuum of information at a time of deep uncertainty and fear – a perfect storm for disinformation.

The pandemic and aptly named *"infodemic"* was exploited by state actors, especially Russia and China, to advance their positions on the global stage. China, in particular, increased its use of aggressive influence operations, coordinating messaging between state media, official social media accounts, and inauthentic networks to target government officials and critics across the world in a bid to improve Beijing's image following its mishandling of the outbreak. COVID-19 mis/disinformation also combined with election-related mis/disinformation, straining the growing community of stakeholders seeking to counter and mitigate the deluge of inaccurate or misrepresented information, particularly that pushed by state actors globally.

Yet, the COVID-19 pandemic also offered an opportunity to test platforms' resolve in addressing inaccurate and harmful information that threatens safety and security. The major social media companies have broadly underplayed their responses to disinformation on their platforms, claiming a commitment to free speech and freedom of expression. Then COVID-19 arrived, and with it, all the mis/disinformation that threatened to make an already devastating situation worse. The major platforms leaped into action, taking steps to aggressively label and flag false and misleading content and outright remove content that spread conspiracy theories and harmful lies. Though far from perfect,

[1] Christina Nemr and William Gangware, "Weapons of Mass Distraction: Foreign State-Sponsored Disinformation in the Digital Age," Park Advisors, March 2019, https://www.park-advisors.com/disinforeport.

their relatively quick responses beg the question, how will they maintain and apply the proactive urgency they displayed early on in the pandemic to other harmful disinformation moving forward?

The online platforms are not alone in this exercise. As concerns over disinformation continue to grow, governments have set up agencies and task forces and taken actions to safeguard their elections and democratic processes from malign foreign influence. Civil society actors globally continue to take proactive measures to mitigate vulnerabilities and strengthen resilience to disinformation as well as reactive measures to analyze, verify, and debunk specific narratives. Technology companies are developing innovative tools to help users uncover disinformation trends and identify particular actors and networks spreading disinformation. This report highlights these efforts and more to increase general understanding of the challenges, the players, the tools, and the continuously evolving responses. It begins with the challenges.

The emergence of COVID-19 in 2020 introduced not only a devastating virus but a **vacuum of information at a time of deep uncertainty and fear – a perfect storm for disinformation.**

2. A Look at Foreign State-Sponsored Disinformation and Propaganda

As the adoption of new technology and social media platforms have increased globally, so too have government efforts to exploit these platforms for their own interests, both at home and abroad. In 2020, disinformation campaigns were reported in 81 countries, up from 70 the year prior (and up from 28 countries in 2018).[2] This includes an increase in disinformation activity emanating from state-sponsored or state-linked entities, for example, through digital ministries, military campaigns, or police force actions.[3] Researchers also observed notable trends in both the objectives and tactics of foreign influence operations, including an increasing use of commercial bot networks and marketing firms, the targeting of multiple countries at once, and greater meddling in several African countries, particularly by Russia.

Though the list of countries partaking in such activities is much longer, this report focuses on recent disinformation and propaganda efforts linked to Russia, China, and Iran.

More is More: Russian Influence and Disinformation Campaigns

Deemed the "firehose of falsehood," Russia's disinformation strategy is known for its high number of dissemination outlets and the spreading of partial truths or falsehoods.[4] In targeting foreign countries, Russia's information warfare machine functions like a diverse and interconnected ecosystem of actors, including state-backed media outlets, social media accounts, intelligence agencies, and cybercriminals.[5] Indeed, an August 2020 report by the U.S. Department of State's Global Engagement Center identified five pillars of the Russian disinformation and propaganda ecosystem: (i) official government resources, (ii)

[2] Samantha Bradshaw, Hannah Bailey, and Philip N. Howard, "Industrialized Disinformation, 2020 Global Inventory of Organized Social Media Manipulation," Oxford Internet Institute, January 2021, https://comprop.oii.ox.ac.uk/research/posts/industrialized-disinformation.

[3] Bradshaw, Bailey, and Howard, "Industrialized Disinformation."

[4] Christopher Paul and Miriam Matthews, "The Russian 'Firehose of Falsehood' Propaganda Model: Why It Might Work and Options to Counter It," RAND Corporation, January 2016, https://www.jstor.org/stable/resrep02439.

[5] Alina Polyakova and Spencer Phipps Boyer, "The future of political warfare: Russia, the West, and the coming age of global digital competition, Brookings Institution, March 2018, https://www.brookings.edu/research/the-future-of-political-warfare-russia-the-west-and-the-coming-age-of-global-digital-competition/.

state-funded international television and radio broadcasting, (iii) proxy resources, (iv) social networks, and (v) disinformation in cyberspace.[6] According to the report, this ecosystem allows for varied and overlapping approaches that reinforce each other and reflect both the sources of disinformation and propaganda and the different tactics these channels use.

A growing web of campaigns

In June 2020, the social network analysis company Graphika detailed the tactics used by long-term Russian influence operation "Secondary Infektion."[7] Beginning in January 2014, the operation consisted of nearly 2,500 articles published in seven languages on 300 different platforms to provoke tensions among Moscow's perceived enemies. While initially only focused on the Russian opposition, the operation evolved to target Ukraine, which it portrayed as a failed state. Other narratives included portraying the United States and the North Atlantic Treaty Organization (NATO) as aggressors, Europe as a divided and weak continent, the Russian government as a victim of Western plots, and opponents of the Moscow regime as corrupt or mentally unstable.

In July 2020, the cybersecurity company FireEye further detailed a sophisticated campaign aligned with Russian security interests that has targeted NATO since 2017.[8] Deemed "Ghostwriter," the campaign primarily targeted audiences in Latvia, Lithuania, and Poland with narratives critical of NATO's presence in Eastern Europe. It also leveraged other themes such as anti-U.S. and COVID-19-related narratives as part of a broader anti-NATO agenda. The campaign involved compromising news websites to plant false articles discrediting NATO or directly disseminating those narratives using spoofed email addresses, so they were shown as coming from government and military officials and journalists, for example.[9] These narratives were further amplified on social media by what FireEye suspected were inauthentic personas, though this was not the primary means of dissemination in the campaign.

Throughout 2020, Twitter and Facebook removed numerous accounts, pages, and groups with links to the Russian state. For instance, in June 2020, Twitter removed 1,152 accounts and 3,434,792 tweets affiliated with Current Policy, a media website engaging in state-backed political propaganda within Russia.[10] In September 2020, Facebook removed approximately 300 assets attributed to the Russian security services as part of a larger takedown targeting multiple clusters publishing disinformation in English, Russian, and Arabic, including a small network of English- and Turkish-language assets linked to Russia's Internet Research Agency.[11] The Kremlin's information operations are often aided by the Russian

[6] "Pillars of Russia's Disinformation and Propaganda Ecosystem," U.S. Department of State, Global Engagement Center, August 2020, https://www.state.gov/russias-pillars-of-disinformation-and-propaganda-report/.

[7] Ben Nimmo, Camille Francois, C. Shawn Eib et al., "Secondary Infektion at a Glance," Graphika, 2020, https://secondaryinfektion.org/report/secondary-infektion-at-a-glance/.

[8] Lee Foster, Sam Riddell, David Mainor, and Gabby Roncone, "'Ghostwriter' Influence Campaign: Unknown Actors Leverage Website Compromises and Fabricated Content to Push Narratives Aligned With Russian Security Interests," *FireEye*, July 29, 2020, https://www.fireeye.com/blog/threat-research/2020/07/ghostwriter-influence-campaign.html.

[9] Foster, Riddell, Mainor, and Roncone, "'Ghostwriter' Influence Campaign."

[10] "Dispatches from the June 2020 Twitter Inauthentic Activity Takedown," Stanford Internet Observatory, June 18, 2020, https://cyber.fsi.stanford.edu/io/news/june-2020-twitter-inauthentic-activity-takedown-russia.

[11] Ben Nimmo, Camille Francois, C. Shawn Eib, Léa Ronzaud, and Joseph Carter, "GRU and the Minions," Graphika, September 24, 2020, https://graphika.com/reports/gru-and-the-minions/; and "Disinformation campaign removed by Facebook linked to Russia's Internet Research Agency," DFRLab, September 24, 2020, https://medium.com/dfrlab/disinformation-campaign-removed-by-facebook-linked-to-russias-internet-research-agency-3cbd88d0dad.

state-controlled international television network RT, which ensures that editorial content is consistent with Russian government foreign policy positions, including the promotion of a strong anti-Western, anti-democracy narrative.[12]

Russia has also recently renewed its interest in Africa, with similar goals to China: consolidate control over cyberspace, undermine Western influence in the region, and capture resources.[13] Researchers found that nearly half of all the foreign influence efforts documented in 2019 targeted African countries or regional groups, many of which originated in Russia.[14] Most notably, Russia used social media manipulation and bribery to support a number of different candidates in the run-up to Madagascar's 2018 presidential elections.[15]

Russia has also recently renewed its interest in Africa, with similar goals to China: **consolidate control over cyberspace, undermine Western influence in the region, and capture resources.**

In October 2020, Facebook removed Russian-linked assets targeting Libya, Sudan, Syria, and the Central African Republic (CAR). The network was identified as being linked to Yevgeny Prigozhin, a Russian oligarch with close ties to the Kremlin and known to be connected to the private military company, the Wagner Group.[16] In Libya, Sudan, and Syria, the network reached an aggregated 5.7 million users.[17] Country-specific narratives aimed to frame Russian activity in the region in a positive light, provoke hostility toward Western operations, and support leaders and groups backed by the Kremlin, such as Syria's Bashar al-Assad. In the CAR, the operation posted primarily about local politics and elections and praised Russia's engagement in the country.[18]

Despite these efforts, the actual impact of Russian influence and disinformation operations is questionable. These operations represent only a small fraction of overall social media engagement, and in most cases, they are unwittingly amplified by media outlets and politicians.[19] Additionally, a recent NATO study on Russian activities in Africa's information environment found that Russia tends to exaggerate its limited influence across the continent; its operations have little impact due to a lack of experience with the environment and access to economic and political influence.[20]

[12] Mona Elswah and Philip N. Howard, "'Anything that causes chaos is RT's line' – new study lifts the lid on RT's role in wreaking political havoc," Oxford Internet Institute, September 28, 2020, https://www.oii.ox.ac.uk/news/releases/anything-that-causes-chaos-is-rts-line-new-study-lifts-the-lid-on-rts-role-in-wreaking-political-havoc/.

[13] Eleonore Pauwels, "The Anatomy of Information Disorders in Africa," *Konrad-Adenauer-Stiftung*, September 9, 2020, https://www.kas.de/en/web/newyork/single-title/-/content/the-anatomy-of-information-disorders-in-africa.

[14] Diego A. Martin, Jacob N. Shapiro, and Julia Ilhardt, "Trends in Online Influence Efforts," Empirical Studies of Conflict Project, 2020, https://esoc.princeton.edu/publications/trends-online-influence-efforts.

[15] Joshua Yaffa, "Is Russian Meddling as Dangerous as We Think?" *The New Yorker*, September 14, 2020, https://www.newyorker.com/magazine/2020/09/14/is-russian-meddling-as-dangerous-as-we-think.

[16] Kimberly Martin, "Exposing and Demanding Accountability for Kremlin Crimes Abroad," Hearing Before the Subcommittee on Europe, Eurasia, Energy, and the Environment, United States House of Representatives, 115th Congress, 2020.

[17] "Stoking Conflict by Keystroke," Stanford Internet Observatory, December 15, 2020, https://cyber.fsi.stanford.edu/io/news/africa-takedown-december-2020.

[18] Nathaniel Gleicher and David Agranovich, "Removing Coordinated Inauthentic Behavior from France and Russia," Facebook, December 15, 2020, https://about.fb.com/news/2020/12/removing-coordinated-inauthentic-behavior-france-russia/.

[19] Joshua Yaffa, "Is Russian Meddling as Dangerous as We Think?" *The New Yorker*, September 14, 2020, https://www.newyorker.com/magazine/2020/09/14/is-russian-meddling-as-dangerous-as-we-think.

[20] "Russia's Activities in Africa's Information Environment," NATO Strategic Communications Center of Excellence, March 2021, https://www.stratcomcoe.org/russias-activities-africas-information-environment-case-studies-mali-central-african-republic.

Russian interference in the U.S. presidential election remained a concern throughout 2020. However, more robust detection and moderation efforts, supported by a whole of society effort since 2016, helped limit the impact of these campaigns. As a result, Russia-linked actors opted for more discreet tactics and strategies, relying on domestic actors to launder narratives, with increased presence on fringe platforms.[21]

In March 2021, U.S. Intelligence services confirmed that Russia presented the primary state-sponsored threat in influencing the 2020 electoral process.[22]

Of course, Russia's arsenal also included active campaigning around the COVID-19 outbreak, using similar tactics to spread false information related to the virus and Western vaccines, undermine Western democracies, and strengthen its influence.[23] For example, Zignal Labs identified a coordinated campaign spreading false COVID-19 information to sow distrust in the Ukrainian government, aiming to pressure the Ukrainian president to try and gain concessions in the Donbas region.[24] The European Union (EU) regularly identifies instances where Russian sources push disinformation and narratives claiming that the outbreak will lead to a dissolution of the EU and NATO.[25] Regarding vaccines specifically, Russian efforts aim to push negative claims about the ones developed by the United States while spreading positive narratives about the Russian Sputnik V vaccine, especially in Africa and Latin America.[26] Russian intelligence services are also taking a more direct role in spreading disinformation around COVID-19, using tactics refined since 2016 that are increasingly difficult to detect, such as the information laundering tactics used in the "Ghostwriter" campaign.[27]

Disarm, Deny, Deflect: Chinese Influence and Disinformation Campaigns

While Russia's present-day campaigns tend to focus on the cyber domain, China's international influence operations are widespread and include economic, political, and personal relationship-building.[28] During the past two decades, the Chinese Communist Party (CCP) expanded its influence over

[21] Maria Snegovaya and Johei Watanabe, "The Kremlin's Social Media Influence Inside the United States: A Moving Target," Free Russia Foundation, February 10 2021, https://www.4freerussia.org/the-kremlin-s-social-media-influence-inside-the-united-states-a-moving-target.

[22] National Intelligence Council, "Foreign Threats to the 2020 US Federal Elections," March 10, 2021, https://www.dni.gov/files/ODNI/documents/assessments/ICA-declass-16MAR21.pdf.

[23] "Disinformation Mash-Up: MH17 and Coronavirus," EUvsDisinfo, March 12, 2020, https://euvsdisinfo.eu/disinformation-mash-up-mh17-and-coronavirus/; and "Analysis | Six reasons the Kremlin spreads disinformation about the coronavirus," DFRLab, March 24, 2020, https://medium.com/dfrlab/commentary-six-reasons-the-kremlin-spreads-disinformation-about-the-coronavirus-8fee41444f60.

[24] Heather McCormish, "The Infodemic of COVID-19: Viral Influence Competition," Zignal Labs, March 19, 2020, https://zignallabs.com/blog/zignal_report/the-infodemic-of-covid-19viral-influence-competition/.

[25] Rikard Jozwiak, "EU Monitors See Coordinated COVID-19 Disinformation Effort By Iran, Russia, China," *Radio Free Europe*, April 22, 2020, https://www.rferl.org/a/eu-monitors-sees-coordinated-covid-19-disinformation-effort-by-iran-russia-china/30570938.html.

[26] "How pro-Kremlin outlets and blogs undermine trust in foreign-made COVID vaccines," DFRLab, January 27, 2021, https://medium.com/dfrlab/how-pro-kremlin-outlets-and-blogs-undermine-trust-in-foreign-made-covid-vaccines-4fa9f9f19df1; "Pro-Kremlin Media as a Marketing Tool: Promoting Sputnik V Vaccine in Latin American," EUvsDisinfo, December 4, 2020, https://euvsdisinfo.eu/pro-kremlin-media-as-a-marketing-tool-promoting-sputnik-v-vaccine-in-latin-america; and John Campbell, "Russian Disinformation popularizes Sputnik V Vaccine in Africa," Council on Foreign Relations, December 10, 2020, https://www.cfr.org/blog/russian-disinformation-popularizes-sputnik-v-vaccine-africa.

[27] Julian E. Barnes and David E. Sanger, "Russian Intelligence Agencies Push Disinformation on Pandemic," *The New York Times*, July 28, 2020, https://www.nytimes.com/2020/07/28/us/politics/russia-disinformation-coronavirus.html.

[28] Abigail Grace, "China's Influence Operations Are Pinpointing America's Weaknesses," *Foreign Policy*, October 4, 2018, https://foreignpolicy.com/2018/10/04/chinas-influence-operations-are-pinpointing-americas-weaknesses/.

media production and dissemination channels around the world, employing more aggressive and technologically sophisticated coercion methods.[29] It also uses varied tactics, including engagement with diaspora communities, propaganda and disinformation, electoral interference, and cyberattacks.[30] In conjunction with the production and global distribution of pro-Chinese media, the CCP acts to reduce awareness of controversial topics by stifling criticism and activism. The CCP also seeks to influence educational and policy institutions abroad and wield financial influence through aggressive loans and infrastructure investments, most notably through the Belt and Road Initiative.[31] These influence efforts aim to paint the country in a positive light, protect its national sovereignty, and define or modify existing global norms to suit the CCP's needs.[32] All these activities combined are part of what is commonly known as China's "Three Warfares," which comprises a hybrid approach of public opinion warfare, psychological warfare, and legal warfare, and is expected to expand in scope and sophistication in the coming years.[33]

Around April 2020, China shifted to more aggressive disinformation and influence operations, **largely to repair its global image after failing to contain COVID-19.**

Increasing aggression

The consistent goal of China's disinformation and influence operations is to protect the CCP's hold on power. In foreign operations, this manifests most often through disinformation campaigns that aim to muddy understanding of undesirable topics, as most recently witnessed with the genocide in Xinjiang, the political repression in Hong Kong, and global recognition of Taiwan as a sovereign nation.[34]

Recently, China's influence operations have expanded. An analysis of Chinese state influence activities on Twitter and Facebook from January 2018 to April 2020 found they primarily targeted Chinese-speaking audiences outside of the Chinese mainland (where Twitter and Facebook are blocked), intending to influence perceptions on key issues, including the Hong Kong protests, exiled Chinese billionaire Guo Wengui, and, to a lesser extent, COVID-19 and Taiwan.[35] Researchers estimated that since the Hong Kong protests in March 2019, the activation of Twitter accounts connected to Chinese embassies, consulates, and ambassadors

[29] Sarah Cook, "Beijing's Global Megaphone," Freedom House, 2020, https://freedomhouse.org/report/special-report/2020/beijings-global-megaphone.

[30] J. Michael Cole, "Exploring China's Political Warfare Against Taiwan," MacDonald Laurier Institute, June 29, 2020, https://www.macdonaldlaurier.ca/exploring-chinas-political-warfare-taiwan-new-paper-j-michael-cole/.

[31] Samantha Custer, Brooke Russell, Matthew DiLorenzo, Mengfan Cheng, Siddhartha Ghose, Harsh Desai, Jacob Sims, and Jennifer Turner, "Ties That Bind: Quantifying China's Public Diplomacy and Its 'Good Neighbor' Effect," Aid Data and Asia Society Policy Institute, June 27, 2018, https://asiasociety.org/policy-institute/ties-bind-quantifying-chinas-public-diplomacy-and-its-good-neighbor-effect.

[32] Dexter Roberts, "China's Disinformation Strategy: Its Dimensions and Future," Indian Strategic Studies, January 1, 2021, http://strategicstudyindia.blogspot.com/2021/01/chinas-disinformation-strategy-its.html.

[33] Peter Mattis, "China's 'Three Warfares' in Perspective," *War on the Rocks*, January 30, 2018, https://warontherocks.com/2018/01/chinas-three-warfares-perspective/.

[34] "Countering Chinese disinformation reports," Atlantic Council, December 17, 2020, https://www.atlanticcouncil.org/in-depth-research-reports/dfrlab-china-reports; Nick Monaco, Melanie Smith, and Amy Studdart, "Detecting Digital Fingerprints: Tracing Chinese Disinformation in Taiwan," Institute for the Future, Graphika, and The International Republican Institute, August 25, 2020, https://www.iftf.org/disinfo-in-taiwan/; and Kate Conger, "Facebook and Twitter Say China Is Spreading Disinformation in Hong Kong," *The New York Times*, August 19, 2020, https://www.nytimes.com/2019/08/19/technology/hong-kong-protests-china-disinformation-facebook-twitter.html.

[35] Jacob Wallis, Tom Uren, Elise Thomas, Albert Zhang, Samantha Hoffman, Lin Li, Alexandra Pascoe, and Danielle Cave, "Retweeting through the Great Firewall: A persistent and undeterred threat actor," Australian Strategic Policy Institute and International Cyber Policy Centre, June 12, 2020, https://www.aspi.org.au/report/retweeting-through-great-firewall.

increased by more than 250 percent.[36] China has also significantly increased disinformation efforts aimed at concealing the mass detention and repression of ethnic minorities in Xinjiang province.[37]

Around April 2020, China shifted to more aggressive disinformation and influence operations, largely to repair its global image after failing to contain COVID-19, with Chinese government officials significantly increasing their social media presence.[38] This increased confrontational approach to manipulate information, seemingly modeled on Russian tactics, included amplifying messaging from Russian and Iranian propaganda outlets, in addition to deflecting responsibility for their role in the spread of COVID-19.[39]

Within Europe, for instance, the Chinese government employed a combination of diplomatic pressure and online influence operations to elevate coverage of its deployment of aid and medical supplies to European countries, a practice referred to as "mask diplomacy."[40] Italy, for example, was a primary target for Chinese influence operations as it had joined the Belt and Road Initiative in 2019, paving the way for Beijing to push its propaganda and use mask diplomacy during the early stages of the pandemic.[41] Chinese diplomats also increasingly used social media platforms like Facebook and Twitter to engage in "wolf warrior" diplomacy, with the ostensible aim to broadcast directly to citizen populations in an effort to contest negative perceptions of the Chinese government or besmirch perceptions of the United States.[42] On Twitter, China has spread disinformation about the origins of COVID-19 and Western government efforts to fight the virus, including promoting the conspiracy theory that Americans brought it to Wuhan.[43]

These operations signaled China's efforts to scale its influence operations further and to experiment with new techniques while doing so.[44] China's military, for example, has expressed interest in using

[36] Jessica Brandt and Bret Schafer, "Five Things to Know About Beijing's Disinformation Approach," Alliance for Securing Democracy, March 30, 2020, https://securingdemocracy.gmfus.org/five-things-to-know-about-beijings-disinformation-approach.

[37] Carmen Molina Acosta, "'Huge uptick' in Chinese propaganda over Uighur camps, report finds," *International Consortium of Investigative Journalists*, July 30, 2020, https://www.icij.org/investigations/china-cables/huge-uptick-in-chinese-propaganda-over-uighur-camps-report-finds.

[38] Sarah Cook, "Welcome to the New Era of Chinese Government Disinformation," *The Diplomat*, May 11, 2020, https://thediplomat.com/2020/05/welcome-to-the-new-era-of-chinese-government-disinformation; and Jessica Brandt and Torrey Taussig, "The Kremlin's disinformation playbook goes to Beijing," Brookings Institution, May 19, 2020, https://www.brookings.edu/blog/order-from-chaos/2020/05/19/the-kremlins-disinformation-playbook-goes-to-beijing/.

[39] Daniel Kliman, Andrea Kendall-Taylor, Kristine Lee, Joshua Fitt, and Carisa Nietsche, "Dangerous Synergies: Countering Chinese and Russian Digital Influence Operations," Center for a New American Security, May 7, 2020, https://www.cnas.org/publications/reports/dangerous-synergies.

[40] Brian Wong, "China's Mask Diplomacy," *The Diplomat*, March 25, 2020, https://thediplomat.com/2020/03/chinas-mask-diplomacy.

[41] Valbona Zeneli and Federica Santoro, "China's Disinformation Campaign in Italy," *The Diplomat*, June 9, 2020, https://thediplomat.com/2020/06/chinas-disinformation-campaign-in-italy/.

[42] Ivana Karásková, Alicja Bachulska, Tamás Matura, Filip Šebok, Matej Šimalčík, "China's Propaganda and Disinformation Campaigns in Central Europe," *MapInfluenCE*, August 2020, https://mapinfluence.eu/en/chinas-propaganda-and-disinformation-campaigns-in-central-europe/; and Zhiqun Zhu, "Interpreting China's Wolf Warrior Diplomacy," *The Diplomat*, May 15, 2020, https://thediplomat.com/2020/05/interpreting-chinas-wolf-warrior-diplomacy/.

[43] Jane Lytvynenko, "Chinese State Media Spread A False Image Of A Hospital For Coronavirus Patients In Wuhan," *BuzzFeed*, January 27, 2020, https://www.buzzfeednews.com/article/janelytvynenko/china-state-media-false-coronavirus-hospital-image; and Ben Westcott and Steven Jiang, "Chinese diplomat promotes conspiracy theory that US military brought coronavirus to Wuhan," *CNN*, March 13, 2020, https://www.cnn.com/2020/03/13/asia/china-coronavirus-us-lijian-zhao-intl-hnk/index.html.

[44] Jacob Wallis, Tom Uren, Elise Thomas, Albert Zhang, Samantha Hoffman, Lin Li, Alexandra Pascoe, and Danielle Cave, "Retweeting through the Great Firewall: A persistent and undeterred threat actor," Australian Strategic Policy Institute and International Cyber Policy Centre, June 12, 2020, https://www.aspi.org.au/report/retweeting-through-great-firewall.

artificial intelligence (AI) to better tailor its messages to influence social media users in Hong Kong, Taiwan, and the United States.[45]

While the effectiveness of its overt COVID-19 influence efforts is debatable, China's broader influence operations have been relatively effective in reshaping some international norms and practices, such as promoting development initiatives free from human rights and other conditions, and, along with Russia, propagating the concept of cyber sovereignty.[46] Referred to as the "Great Firewall," China's vision of state sovereignty over internet access and content is attractive to governments seeking to impose greater online surveillance, censorship, and other digital authoritarian tools to stifle opposition and control their populations. However, as China expands its use of disinformation alongside economic and political instruments of influence, there is growing awareness of the CCP's strategic intent and the adverse effects its tactics have on democratic institutions.[47]

Iranian disinformation and influence operations have displayed similarities and, in some cases, **even convergence with Russian and Chinese tactics.**

Local Priorities, Some Global Forays: Iranian Influence and Disinformation Campaigns

Iranian disinformation and influence operations have displayed similarities and, in some cases, even convergence with Russian and Chinese tactics. Like Russia, Iran's disinformation efforts typically oppose U.S. foreign policy, particularly in the Middle East and North Africa (MENA), and take advantage of contentious domestic issues in the United States.[48]

In the MENA region, Iranian and Russian trolls coordinated their efforts to obscure responsibility for violence by the Syrian government and pushed narratives favorable to the Syrian Armed Forces.[49] In April 2020, Facebook removed more than 500 pages, groups, and accounts attributed to the Islamic Republic of Iran Broadcasting Corporation.[50] The takedown covered nine years (2011-2020) of activity conducted in multiple languages, across four continents (Africa, Asia, Europe, and North America), that posted about a wide range of themes, including anti-Israel and anti-Saudi Arabia messaging.[51] In addition, Hizballah reportedly built troll farms across the Middle East to amplify pro-Iranian disinformation campaigns, which included training on how to digitally manipulate photographs, manage large numbers of fake social media accounts, and avoid Facebook's censorship.[52]

[45] "China's military aims to use AI to dominate in cyber and outer space, Japanese think tank warns," *South China Morning Post*, November 13, 2020, https://www.scmp.com/news/china/military/article/3109803/chinas-military-aims-use-ai-dominate-cyber-and-outer-space.

[46] Matt Schrader, "Friends and Enemies: A Framework for Understanding Chinese Political Interference in Democratic Countries," Alliance for Securing Democracy, April 22, 2020, https://securingdemocracy.gmfus.org/friends-and-enemies-a-framework-for-understanding-chinese-political-interference-in-democratic-countries/.

[47] "A World Safe for the Party: China's Authoritarian Influence and the Democratic Response," International Republican Institute, February 2021, https://www.iri.org/resource/china-expands-global-authoritarian-influence-efforts-some-fragile-democracies-show.

[48] Clint Watts, "Triad of Disinformation: How Russia, Iran, & China Ally in a Messaging War against America," Alliance for Securing Democracy, May 15, 2020, https://securingdemocracy.gmfus.org/triad-of-disinformation-how-russia-iran-china-ally-in-a-messaging-war-against-america/.

[49] Watts, "Triad of Disinformation."

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

[51] Ben Nimmo, C. Shawn Eib, Léa Ronzaud, Rodrigo Ferreira, Thomas Lederer, and Melanie Smith, "Iran's Broadcaster: Inauthentic Behavior," Graphika, May 5, 2020, https://graphika.com/reports/irans-broadcaster-inauthentic-behavior.

[52] Wil Crisp and Suadad al-Salhy, "Exclusive: Inside Hizbollah's fake news training camps sowing instability across the Middle East," *The Telegraph*, August 2, 2020, https://www.telegraph.co.uk/news/2020/08/02/exclusive-inside-hezbollahs-fake-news-training-camps-sowing.

Furthermore, along with Russia and China, Iran has targeted its disinformation operations at foreign audiences, exploiting COVID-19 and foreign protest movements to push its foreign policy agenda.[53] For example, Iranian government officials, journalists, and personalities parroted and amplified China's disinformation about the virus being created by the United States.[54] Former Iranian President Mahmoud Ahmadinejad went as far as sending a conspiracy-laden letter to the World Health Organization claiming the virus is a biological weapon created in unidentified laboratories.[55] A prominent Iranian influence actor, the International Union of Virtual Media (IUVM), produced, disseminated, and amplified narratives blaming the United States and praising China's role in responding to COVID-19.[56] The IUVM's articles on Western press coverage also accused the media of propaganda and psychological operations against China and Iran. Prominent Iranian government and state media accounts also amplified an anti-Israel hashtag campaign related to the virus, including through grassroots Shia organizations in Nigeria and Pakistan and a cluster of similar Iranian accounts posting in a coordinated manner.[57]

In October 2020, Twitter and Facebook removed a tranche of accounts, groups, and pages linked to Iran for violating their platforms' policies on state-linked information operations or coordinated inauthentic behavior.[58] However, these campaigns were not widespread, and their unrefined and sporadic messaging, ineffective dissemination, and inability to build a loyal audience base limited their impact. Furthermore, in October 2020, the U.S. Justice Department (DOJ) seized 92 websites used by Iran's Islamic Revolutionary Guard Corps (IRGC) to spread disinformation and influence campaigns. Four websites were identified as being based in the United States, while 88 were based in Europe, the Middle East, and Southeast Asia.[59] In November 2020, DOJ seized an additional 27 domains used by the IRGC for breaching the Foreign Agents Registration Act.[60] Among those sanctioned was the Islamic Radio and Television Union, which operates as a network of more than 200 media outlets in 35 countries, including in Western Europe and the United States. This network aims to influence sociopolitical events

[53] "The Alliance for Securing Democracy Expands Hamilton 2.0 Dashboard to Include Iran," *Alliance for Securing Democracy*, June 11, 2020, https://securingdemocracy.gmfus.org/the-alliance-for-securing-democracy-expands-hamilton-2-0-dashboard-to-include-iran/.

[54] Michael Lipin, Liyuan Lu, Behrooz Samadbeygi, and Mehdi Jedinia, "Iran, China Amplify Each Other's Allegations of US Coronavirus Culpability," *Voice of America*, March 24, 2020, https://www.voanews.com/middle-east/voa-news-iran/iran-china-amplify-each-others-allegations-us-coronavirus-culpability.

[55] David Brennan, "Former Iranian President Spreads Coronavirus Conspiracy Theory, Calls on World Health Organization to Identify 'Perpetrators' of 'Biological War'," *Newsweek*, March 11, 2020, https://www.newsweek.com/former-iranian-president-spreads-coronavirus-conspiracy-theory-calls-world-health-organization-1491644.

[56] Ben Nimmo, Camille François, C. Shawn Eib, and Léa Ronzaud, "Iran's IUVM Turns to Coronavirus," Graphika, April 2020, https://graphika.com/reports/irans-iuvm-turns-to-coronavirus/.

[57] Abuzar Royesh and Shelby Grossman, "#Covid1948: The Spread of an Anti-Israel Hashtag," Stanford Internet Observatory, August 13, 2020, https://cyber.fsi.stanford.edu/io/news/covid1948-hashtag.

[58] "Disclosing networks to our state-linked information operations archive," Twitter, October 8, 2020, https://blog.twitter.com/en_us/topics/company/2020/disclosing-removed-networks-to-our-archive-of-state-linked-information.html/; and "October 2020 Coordinated Inauthentic Behavior Report," Facebook, November 5, 2020, https://about.fb.com/news/2020/11/october-2020-cib-report. For analyses of these takedowns, see: Carly Miller, Tara Kheradpir, Renee DiResta, and Abuzar Royesh, "Hacked and Hoaxed: Tactics of an Iran-Linked Operation to Influence Black Lives Matter Narratives on Twitter," Stanford Internet Observatory, October 8, 2020, https://cyber.fsi.stanford.edu/io/news/twitter-takedown-iran-october-2020; and "Analysis of an October 2020 Facebook Takedown Linked to the Islamic Movement in Nigeria," Stanford Internet Observatory, October 8, 2020, https://cyber.fsi.stanford.edu/news/islamic-movement-nigeria.

[59] Kartikay Mehrotra, "U.S. Seizes 92 Websites Used by Iran to Spread Disinformation," *Bloomberg*, October 7, 2020, https://www.bloomberg.com/news/articles/2020-10-08/u-s-seizes-92-websites-used-by-iran-to-spread-disinformation?sref=SXpDZ9y7.

[60] Campbell Kwan, "US seizes another crop of Iranian propaganda domains masked as news outlets," *ZDNet*, November 5, 2020, https://www.zdnet.com/article/us-seizes-another-crop-of-iranian-propaganda-domains-masked-as-news-outlets.

to Iran's advantage, such as government responses to COVID-19, and uses false content on social media platforms and state-run and -backed media outlets.[61]

Despite some similarities to and convergences with Russia and China's tactics, Iran's efforts are not as sophisticated, and there is less evidence of coordination among Iran's different campaigns.[62] Its online influence operations are often sloppy with misspelled names; it uses recycled tactics and techniques and lacks familiarity with target audiences and their behavior.[63] Recognizing its limitations, Iran focuses its efforts regionally, investing in Arabic-language media outlets and amplifying social media content favorable to its interests.[64]

[61] "Spotlight: Kharon - Investigating Iran's influence and disinformation operations," Disinfo Cloud, November 16, 2020, https://disinfocloud.com/blog/spotlight-kharon.

[62] Diego A. Martin, Jacob N. Shapiro, and Julia Ilhardt, "Trends in Online Influence Efforts," Empirical Studies of Conflict Project, 2020, https://esoc.princeton.edu/publications/trends-online-influence-efforts.

[63] Ariane Tabatabai, "Iran's Authoritarian Playbook: The Tactics, Doctrine, and Objectives behind Iran's Influence Operations," Alliance for Securing Democracy, September 17, 2020, https://securingdemocracy.gmfus.org/irans-authoritarian-playbook.

[64] Tabatabai, "Iran's Authoritarian Playbook."

3. Who's Responding and How?

Government-Led Responses to Disinformation

The rise of task forces

As concerns over foreign influence operations continue to grow, countries have set up agencies and task forces to address foreign state-sponsored disinformation and propaganda. Germany, a primary target for Russian disinformation, created a unit within the Foreign Ministry's Office for Strategic Communications to counter malign influence operations and disinformation.[65] In April 2017, the government also established a dedicated Cyber and Information Domain Service to protect German citizens from cyber and information-related threats, including foreign influence operations.[66] The Australian government, partly prompted by rising tensions with Beijing, exacerbated during the COVID-19 pandemic, announced its intent in 2020 to establish a task force within the Department of Foreign Affairs and Trade to combat manipulative online content designed to sow internal confusion and division.[67] This task force builds on earlier efforts to counter disinformation. In April 2019, the Australian government launched a public awareness campaign entitled "Stop and Consider" to encourage voters to pay attention to their information sources ahead of the May 2019 federal elections.[68]

Some countries also took specific actions to safeguard their elections and democratic processes from malign foreign influence. Amid warnings of potential foreign interference, the Australian government set

[65] "Außenpolitik strategisch kommunizieren - Werte und Interessen gezielter vermitteln," *Auswärtiges Amt*, May 25, 2018, https://www.auswaertiges-amt.de/de/aussenpolitik/themen/-/2089138.

[66] Sumi Somaskanda, "Germany struggles to step up cyberdefense," *DW*, August 7, 2018, https://www.dw.com/en/germany-struggles-to-step-up-cyberdefense/a-44979677.

[67] Grant Wyeth, "Australia Aims to Combat Disinformation," *The Diplomat*, June 17, 2020, https://thediplomat.com/2020/06/australia-aims-to-combat-disinformation/; Natasha Kassam, "Great expectations: The unraveling of the Australia-China relationship," Brookings Institution, July 20, 2020, https://www.brookings.edu/articles/great-expectations-the-unraveling-of-the-australia-china-relationship/; Sarah Morrison, Belinda Barnet, and James Martin, "China's disinformation threat is real. We need better defences against state-based cyber campaigns," *The Conversation*, June 20, 2020, https://theconversation.com/chinas-disinformation-threat-is-real-we-need-better-defences-against-state-based-cyber-campaigns-141044; and Stephen Dziedzic and Melissa Clarke, "Morrison Government plans to set up taskforce to counter online disinformation," *ABC News*, June 16, 2020, https://www.abc.net.au/news/2020-06-17/foreign-minister-steps-up-criticism-china-global-cooperation/12362076.

[68] "AEC encouraging voters to 'stop and consider' this federal election," Australian Electoral Commission, April 15, 2019, https://www.aec.gov.au/media/media-releases/2019/04-15.htm.

up the Electoral Integrity Assurance (AEC) task force in June 2018, led by the Home Affairs Department, to identify potential cyberattacks and foreign influence campaigns targeting national elections.[69] Alongside Microsoft and the Alliance for Securing Democracy, Canada will lead a global effort to counter the use of cyberattacks and disinformation campaigns to disrupt elections.[70] Other efforts involve engaging election authorities and digital platforms to ensure a well-informed electorate by, for example, identifying and blocking false information, providing fact-checking resources to the general public, and amplifying credible information during election season and beyond. The Canadian government called on social media platforms to do more to combat disinformation ahead of the 2019 election and pressured technology companies to be more transparent about their anti-disinformation and advertising policies.[71]

Improving societal resilience and mitigating the effects of disinformation is also a key component of some government-led responses.

Nationally-led, citizen-focused

Government-led approaches also include efforts to understand the dynamics behind disinformation narratives and how to deploy countermeasures through policy responses and national and international counter-disinformation campaigns, for example.[72] In November 2020, the Spanish government approved measures that allow the government to carry out counter-disinformation campaigns and define what constitutes "misinformation," while working in collaboration with the EU.[73] While opposition parties criticized the new initiative for encroaching on press freedom, it received support from the European Commission, which deemed it an important step in guaranteeing Spain's participation in the EU's Action Plan Against Disinformation.[74] In addition to a government website dedicated to answering citizens' questions about the COVID-19 vaccine, Spain's health ministry is launching a WhatsApp interactive channel to fight vaccine-related disinformation.[75]

Improving societal resilience and mitigating the effects of disinformation is also a key component of some government-led responses. Finland has been at the forefront of cultivating media literacy as an effective counter-disinformation measure, having implemented a program in schools to teach students how to read news critically and equip them with a digital literacy toolkit.[76] To help promote critical thinking, the program emphasizes universal values upheld by Finnish society, including fairness,

[69] M. Moon, "Australia task force will protect elections against cyberattacks," *Engadget*, June 10, 2018, https://www.engadget.com/2018-06-10-australia-task-force-elections-cyberattack.html; and "Electoral Integrity Assurance Taskforce," *Australian Electoral Commission*, last updated July 10, 2020, https://www.aec.gov.au/elections/electoral-advertising/electoral-integrity.htm.

[70] Maggie Miller, "Canada to lead global effort to counter election interference," *The Hill*, May 26, 2020, https://thehill.com/policy/cybersecurity/499598-canada-to-lead-global-effort-to-counter-election-interference.

[71] "Statements and Speeches," Elections Canada, accessed on January 14, 2021, https://www.elections.ca/content.aspx?section=med&dir=spe&document=nov0618&lang=e.

[72] Kalina Bontcheva and Julie Posetti, "Balancing Act: Countering Digital Disinformation while respecting Freedom of Expression," International Telecommunication Union and the UN Educational, Scientific and CulturalOrganization (UNESCO), 2020, https://en.unesco.org/publications/balanceact.

[73] Miguel González and Natalia Junquera, "Spain to monitor online fake news and give a 'political response' to disinformation campaigns," *El Pais*, November 9, 2020, https://english.elpais.com/politics/2020-11-09/spain-to-monitor-online-fake-news-and-give-a-political-response-to-disinformation-campaigns.html.

[74] Bernardo de Miguel, "EU Commission backs Spain's protocol against disinformation campaigns," *El Pais*, November 10, 2020, https://english.elpais.com/politics/2020-11-10/eu-commission-backs-spains-protocol-against-disinformation-campaigns.html.

[75] Fernando Heller, "Spain to launch Whatsapp channel to fight vaccine disinformation," *EurActiv*, January 13, 2021, https://www.euractiv.com/section/digital/news/spain-to-launch-whatsapp-channel-to-fight-vaccine-disinformation/.

[76] Jon Henley, "How Finland starts its fight against fake news in primary schools, *The Guardian*, January 29, 2020, https://www.theguardian.com/world/2020/jan/28/fact-from-fiction-finlands-new-lessons-in-combating-fake-news.

the rule of law, respect for others' differences, openness, and freedom.[77] Furthermore, students are taught professional fact-checking methods, helping them understand different types of misleading information.[78] Canada's Digital Citizen Initiative (DCI) also aims to build societal resilience to disinformation and establishes partnerships to support a healthy information ecosystem.[79] The DCI has provided more than 10 million dollars in grants and contributions since 2019 to support civic literacy programming and tools, research, and civil society and academic capacity building to better understand and build resilience to online disinformation in Canada.[80]

Legislating the disinformation away

Some countries have taken more heavy-handed legislative actions and measures to regulate both social media and big tech platforms to crack down on illegal or harmful online content. Many of these responses raise concerns regarding their implications for freedom of speech and other human rights violations and, in some cases, have been used to target activists and opposition figures.[81]

Chief among these regulations is Germany's 2018 Network Enforcement Act or NetzDG. As part of a wider push against extremism and hate speech, NetzDG requires online platforms to remove "illegal" posts within 24 hours or face substantial fines.[82] NetzDC was expanded in 2020, requiring platforms to report identified "criminal content" to the Federal Criminal Police Office.[83] While broader reforms to the law are being considered to bolster user rights and transparency reporting requirements for platforms, privacy advocates warn that the law could have damaging long-term effects on Germany's hard-earned data protection standards. It could also potentially require platforms to reveal user identities to authorities without a court order, as is currently the law.[84] As the first European regulation of its kind, NetzDG served as a source of inspiration for other regulators and countries like Singapore.[85]

[77] Thomas Roulet, "To combat conspiracy theories teach critical thinking – and community values," *The Conversation*, November 9, 2020, https://theconversation.com/to-combat-conspiracy-theories-teach-critical-thinking-and-community-values-147314; and Eliza Mackintosh, "Finland is winning the war on fake news. What it's learned may be crucial to Western democracy," *CNN*, May 2019, https://edition.cnn.com/interactive/2019/05/europe/finland-fake-news-intl/.

[78] Emma Charlton, "How Finland is fighting fake news – in the classroom," *World Economic Forum*, May 21, 2019, https://www.weforum.org/agenda/2019/05/how-finland-is-fighting-fake-news-in-the-classroom/; and Markus Neuvonen, Kari Kivinen, and Mikko Salo, "Fact checking for educators and future voters," *FactBar*, 2018.

[79] Government of Canada, "Online Disinformation," modified August 17, 2020, https://www.canada.ca/en/canadian-heritage/services/online-disinformation.html.

[80] "Digital Citizen Initiative," UNESCO, accessed January 8, 2021, https://en.unesco.org/creativity/policy-monitoring-platform/digital-citizen-initiative.

[81] Alana Schetzer, "Governments are making fake news a crime – but it could stifle free speech," *The Conversation*, accessed January 16, 2021, https://theconversation.com/governments-are-making-fake-news-a-crime-but-it-could-stifle-free-speech-117654.

[82] "Overview of the NetzDG Network Enforcement Law," Center for Democracy and Technology, July 17, 2017, https://cdt.org/insights/overview-of-the-netzdg-network-enforcement-law/.

[83] Phillip Grüll, "German online hate speech reform criticised for allowing 'backdoor' data collection," *EurActiv*, June 19, 2020, https://www.euractiv.com/section/data-protection/news/german-online-hate-speech-reform-criticised-for-allowing-backdoor-data-collection/.

[84] Natasha Lomas, "Germany tightens online hate speech rules to make platforms send reports straight to the feds," *TechCrunch*, June 19, 2020, https://techcrunch.com/2020/06/19/germany-tightens-online-hate-speech-rules-to-make-platforms-send-reports-straight-to-the-feds/?guccounter=1; and Janosch Delcker, "Germany's balancing act: Fighting online hate while protecting free speech," *POLITICO*, October 1, 2020, https://www.politico.eu/article/germany-hate-speech-internet-netzdg-controversial-legislation/.

[85] "Germany: Flawed Social Media Law," Human Rights Watch, February 14, 2018, https://www.hrw.org/news/2018/02/14/germany-flawed-social-media-law#.

Indeed, in May 2019, Singapore criminalized the dissemination of false information online that compromises security, public safety, and the country's relations with other nations.[86] Individuals violating the law could face heavy fines and jail time, including a potential 10-year sentence if the falsehood were shared using inauthentic means like a fake or bot account.[87] Platforms like Facebook also face fines of as much as US$740,000 and jail sentences of up to 10 years for their roles in spreading misinformation. The law further outlaws the spread of misinformation on private messaging apps and gives the government broad powers to remove false content that undermines public trust in the government.[88]

Similarly, Kenya's 2018 act criminalizes the publication of false information, imposing a fine of US$50,000 and up to two years in prison for publishing false information intentionally.[89] In March 2020, an individual was arrested for spreading misleading information about COVID-19 in violation of the law.[90]

During the pandemic, the Brazilian government pushed a controversial "Fake News" bill intended to target those that finance disinformation on social media. The bill is arguably one of the world's most restrictive internet laws to civil liberties and freedom. Individuals can face up to five years in prison for creating or sharing content deemed to pose a serious risk to "social peace or economic order." It could also undermine the privacy of communications, which would damage the country's digital rights framework pioneered in 2014.[91]

Additionally, in the absence of robust responses, governments are increasingly turning to arbitrary internet and social media shutdowns to grapple with the surge of false information amplified through social media platforms, especially during crises.[92] Some governments have also used the COVID-19 outbreak as a pretext to implement more authoritative measures to maintain control.[93] Such efforts include criminalizing the promotion of "fake news," expanding existing penalties for spreading misinformation and disinformation and increasing surveillance and censorship.[94] While these shutdowns

[86] Jon Russell, "Singapore passes controversial 'fake news' law which critics fear will stifle free speech," *TechCrunch*, May 9, 2019, https://techcrunch.com/2019/05/09/singapore-fake-news-law/.

[87] Kirsten Han, "Singapore to decide what's fake and what's real," *Asia Times*, October 4, 2019, https://asiatimes.com/2019/10/singapore-to-decide-whats-fake-and-whats-real/.

[88] Tessa Wong, "Singapore fake news law polices chats and online platforms," *BBC*, May 9, 2019, https://www.bbc.com/news/world-asia-48196985; and James Griffiths, "Singapore just used its fake news law. Critics say it's just what they feared," *CNN*, November 30, 2019, https://www.cnn.com/2019/11/29/media/singapore-fake-news-facebook-intl-hnk/index.html.

[89] Dominic Inokhomi and John Syekei, "The Computer Misuse and Cybercrimes Act," Bowman's Law, March 6, 2020, https://www.bowmanslaw.com/insights/finance/the-computer-misuse-and-cybercrimes-act.

[90] Busang Senne, "Kenyan man arrested for spreading fake news on coronavirus," *Times Live*, March 16, 2020, https://www.timeslive.co.za/news/africa/2020-03-16-kenyan-man-arrested-for-spreading-fake-news-on-coronavirus/.

[91] "Brazil: Reject 'Fake News' Bill," Human Rights Watch, June 24, 2020, https://www.hrw.org/news/2020/06/24/brazil-reject-fake-news-bill; Raphael Tsavkko Garcia, "Brazil's 'fake news' bill won't solve its misinformation problem," *Technology Review*, September 10, 2020, https://www.technologyreview.com/2020/09/10/1008254/brazil-fake-news-bill-misinformation-opinion/; and Daniel O'Maley, "How Brazil Crowdsourced a Landmark Law," *Foreign Policy*, January 19, 2016, https://foreignpolicy.com/2016/01/19/how-brazil-crowdsourced-a-landmark-law/.

[92] Dave Lawler and Sara Fischer, "Internet blackouts skyrocket amid global political unrest," *Axios*, February 2, 2021, https://www.axios.com/internet-blackouts-myanmar-global-unrest-c2b310d7-d9c4-42f7-9d17-f712527da3ea.html.

[93] Jenna Hand, "'Fake news' laws, privacy & free speech on trial: Government overreach in the infodemic?" *First Draft*, August 12, 2020, https://firstdraftnews.org/latest/fake-news-laws-privacy-free-speech-on-trial-government-overreach-in-the-infodemic/.

[94] "Censorious governments are abusing 'fake news' laws," *The Economist*, February 11, 2021, https://www.economist.com/international/2021/02/13/censorious-governments-are-abusing-fake-news-laws.

may provide the semblance of short-term relief, they fail to address the root causes of disinformation during crises, as citizens still lack access to authoritative information.[95] Moreover, they inflict various social, economic, and humanitarian harms, all of which have been exacerbated during the pandemic due to increased reliance on the internet to access health care, education, and work.[96] As internet shutdowns grow in use across democratic and authoritarian states alike, they risk being exploited by illiberal leaders to stifle freedom of expression and consolidate power. The Indian government, for example, imposed the most internet shutdowns in 2020, claiming that they were precautionary measures to maintain law and order and curtail the spread of mis/disinformation, but some analysis suggests they provided cover for state violence.[97]

Of all the government-led approaches to counter disinformation, legislation is the most fraught due to ill-defined terms and the potential for sweeping applicability. The shortcomings of such legislation, particularly those previously mentioned, underscore the need for continuous engagement across industries, including with civil society organizations, media, academia, and private sector companies. Cross-industry collaboration, complicated as it may be, is key to building shared rules around technology governance in line with democratic norms that account for the varied experiences of different groups online. The following sections show what the private sector and civil society have been doing on this front and where opportunities for such collaboration may arise.

CASE STUDY: EUROPEAN UNION

In 2018, the European Commission created a non-binding action plan for member states dealing with disinformation in Europe and beyond.[98] It focuses on four pillars: (i) improving the capabilities of EU institutions to detect, analyze, and expose disinformation, (ii) strengthening coordinated and joint responses to disinformation, (iii) mobilizing the private sector to tackle disinformation, and (iv) raising awareness and improving societal resilience.

[95] Jayshree Bajoria, “India Internet Clampdown Will Not Stop Misinformation,” Human Rights Watch, April 24, 2019, https://www.hrw.org/news/2019/04/24/india-internet-clampdown-will-not-stop-misinformation; and Jan Rydzak, “Shutting down social media does not reduce violence, but rather fuels it,” *The World*, April 29, 2019, https://www.pri.org/stories/2019-04-29/shutting-down-social-media-does-not-reduce-violence-rather-fuels-it.

[96] Darrell M. West, “Shutting down the internet,” Brookings Institution, February 5, 2021, https://www.brookings.edu/blog/techtank/2021/02/05/shutting-down-the-internet/; and “Policy Brief: Internet Shutdowns,” The Internet Society, December 18, 2019, https://www.internetsociety.org/policybriefs/internet-shutdowns/.

[97] Hanna Duggal, “Mapping internet shutdowns around the world,” *Al Jazeera*, March 3, 2021, https://www.aljazeera.com/news/2021/3/3/mapping-internet-shutdowns-around-the-world; and Nehal Johri, “India’s internet shutdowns function like ‘invisibility cloaks’,” *DW*, November 13, 2020, https://www.dw.com/en/indias-internet-shutdowns-function-like-invisibility-cloaks/a-55572554.

[98] European Commission, “Action Plan against Disinformation,” December 11, 2018, https://ec.europa.eu/digital-single-market/en/news/action-plan-against-disinformation.

The commission further developed a Code of Practice on Disinformation for social media platforms.[99] Current signatories include Facebook, Twitter, Mozilla, Microsoft, Google, TikTok, and associations and members of the advertising industry. The Code of Practice contains a wide variety of commitments, including increased transparency in political advertising, the closure of fake accounts, and the demonetization of disinformation producers.[100] The Code of Practice helped shape platforms' approach to the COVID-19 infodemic in Europe, for instance, by incentivizing policies to increase the visibility of authoritative information and the removal of misleading and harmful content. Signatories of the Code of Practice have thus far provided five sets of reports on the measures they adopted to counter disinformation during the pandemic.[101] However, EU lawmakers say the Code of Practice fails to provide sufficient incentives for platforms to abide by transparency standards due to inconsistent and incomplete application of the code across platforms and member states, lack of uniform definitions, gaps in the coverage of the code commitments, and other limitations intrinsic to self-regulation.[102]

In December 2020, the European Commission released its European Democracy Action plan, which includes steps to improve the EU's toolbox to counter foreign disinformation. The plan aims to reduce the economic incentives for spreading disinformation and impose costs on actors engaging in state-sponsored influence operations and disinformation campaigns. For example, the EU plans to introduce targeted sanctions on perpetrators following repeated offenses, thereby raising the cost of foreign disinformation operations.[103]

The European Commission outlined efforts to enhance the Code of Practice on Disinformation by spring 2021, so it aligns with the EU Digital Services Act. Platforms will be required to assess the risks their systems pose to the public interest and develop risk management tools in response.[104] Other EU initiatives focus on funding projects targeting misinformation and engaging civil society organizations and higher education institutions.[105] The European Digital Media Observatory, for example, aims to create a collaborative hub for EU-based fact-checkers, academics, and

[99] European Commission, "Code of Practice on Disinformation," last updated March 16, 2021, https://ec.europa.eu/digital-single-market/en/code-practice-disinformation.

[100] European Commission, "Statement by Commissioner Gabriel on the Code of Practice on Online Disinformation," September 26, 2018, https://ec.europa.eu/commission/presscorner/detail/en/STATEMENT_18_5914.

[101] European Commission, "Fifth set of reports – Fighting COVID-19 disinformation Monitoring Programme," January 18, 2021, https://ec.europa.eu/digital-single-market/en/news/fifth-set-reports-fighting-covid-19-disinformation-monitoring-programme.

[102] European Commission, "First baseline reports – Fighting COVID-19 disinformation Monitoring Programme," September 10, 2020, https://ec.europa.eu/digital-single-market/en/news/first-baseline-reports-fighting-covid-19-disinformation-monitoring-programme; and Natasha Lomas, "EU lawmakers say it's time to go further on tackling disinformation," *Tech Crunch*, September 10, 2020, https://techcrunch.com/2020/09/10/eu-lawmakers-say-its-time-to-go-further-on-tackling-disinformation/.

[103] European Commission, "Communication from the Commission on the European Democracy Action Plan," December 3, 2020, https://ec.europa.eu/info/sites/info/files/edap_communication.pdf.

[104] European Commission, "Communication from the Commission on the European Democracy Action Plan," December 3, 2020, https://ec.europa.eu/info/sites/info/files/edap_communication.pdf.

[105] European Commission, "10 ways the EU is fighting disinformation," *Medium*, September 18, 2019, https://europeancommission.medium.com/10-ways-the-eu-is-fighting-disinformation-f07fca60e918; and European Commission, "Funding opportunities about Disinformation," accessed on January 15, 2021, https://ec.europa.eu/digital-single-market/en/newsroom-agenda/funding-opportunity/disinformation.

other relevant stakeholders.[106] The Commission plans to establish new procedures to ensure effective collaboration between fact-checkers and adequate data disclosures for research on disinformation.[107]

CASE STUDY: TAIWAN

The Taiwanese government developed a comprehensive approach to respond to adversarial propaganda, best illustrated during its 2020 presidential election. Taiwan managed to resist growing Chinese disinformation operations through partnerships with civil society groups and robust, consistent communication with the technology industry.[108] Prior to the election, a cross-government communications group was created to monitor social media and analyze online political conversations for propaganda and disinformation.[109] Government agencies were also tasked with putting out clarifying messages, sometimes using memes, to refute false or misleading claims quickly.[110] Furthermore, in the run-up to the 2020 election, the government drove media-literacy trucks to rural areas and conducted workshops to educate citizens on the dangers of disinformation.[111]

To address Chinese disinformation, the Taiwanese government enacted several executive and legislative actions in the lead-up to the 2020 election.[112] Most notably, the 2019 Anti-Infiltration law penalizes the activities of "foreign hostile forces," such as making political donations,

[106] European Commission, "The European Digital Media Observatory," March 16, 2021, https://ec.europa.eu/digital-single-market/en/european-digital-media-observatory.

[107] European Commission, "Communication from the Commission on the European Democracy Action Plan," December 3, 2020, https://ec.europa.eu/info/sites/info/files/edap_communication.pdf.

[108] Nick Monaco, Melanie Smith, and Amy Studdart, "Detecting Digital Fingerprints;: Tracing Chinese Disinformation in Taiwan, Institute for the Future, Graphika, and The International Republican Institute, August 2020, https://www.iftf.org/disinfo-in-taiwan; and Jacob Mchangama and Jonas Parello-Plesner, "Taiwan's Disinformation Solution," *The American Interest*, February 6, 2020, https://www.the-american-interest.com/2020/02/06/taiwans-disinformation-solution/.

[109] Aaron Huang, "Combatting and Defeating Chinese Propaganda and Disinformation: A Case Study of Taiwan's 2020 Elections," Harvard Kennedy School Belfer Center, July 2020, https://www.belfercenter.org/publication/combatting-and-defeating-chinese-propaganda-and-disinformation-case-study-taiwans-2020.

[110] Chun Han Wong and Philip Wen, "Taiwan Turns to Facebook and Viral Memes to Counter China's Disinformation," *Wall Street Journal*, January 3, 2020, https://www.wsj.com/articles/taiwan-turns-to-facebook-and-viral-memes-to-counter-chinas-disinformation-11578047403.

[111] Aaron Huang, "Combatting and Defeating Chinese Propaganda and Disinformation: A Case Study of Taiwan's 2020 Elections," Harvard Kennedy School Belfer Center, July 2020, https://www.belfercenter.org/publication/combatting-and-defeating-chinese-propaganda-and-disinformation-case-study-taiwans-2020.

[112] Linda Zhang, "How to Counter China's Disinformation Campaign in Taiwan," *Military Review*, September-October 2020, https://www.armyupress.army.mil/Journals/Military-Review/English-Edition-Archives/September-October-2020/Zhang-Disinformation-Campaign/.

spreading disinformation, staging campaign events, or otherwise interfering in elections.[113] Under this law, any person or entity receiving such support faces imprisonment of up to five years and fines of more than US$330,000. Additionally, the Ministry of Justice established the Big Data and Public Opinion Task Force, and security institutions, including the Ministry of National Defense and the National Security Council, coordinated responses to Chinese disinformation.[114] The government also enacted laws to punish traditional media outlets suspected of promoting narratives from mainland China and banned select Chinese media platforms like QIYI (Baidu's video platform) and Tencent video from the Taiwanese market, though they continue to operate through legal loopholes.[115]

Cooperation with civil society and the civic tech community, such as g0v, enabled the Taiwanese government to collect and analyze vast amounts of data and carefully target its response to maximize efficiency and reach.[116] Civil society has been particularly active in identifying, tracking, and flagging disinformation through a number of civic tech initiatives like DoubleThink Labs, the Open Culture Foundation, and CoFacts.[117] These actions, combined with strong public engagement through educational programs and high levels of transparency, have helped inform and enlist public support.[118] The secondary effects of Taiwan's whole-of-society approach are seen with its successful containment of COVID-19.[119] Crowdsourced fact-checking groups flagged stories around COVID-19's spread early in the pandemic and supported the government in developing a rapid response.[120] Based on its relative success with countering disinformation, Taiwan has collaborated with countries around the world to share its insights and best practices.

[113] KG Chan, "Taiwan's new anti-infiltration law aimed at China," *Asia Times*, December 31, 2020, https://asiatimes.com/2019/12/taiwans-new-anti-infiltration-law-aimed-at-china/.

[114] Yuki Tatsumi, Pamela Kennedy, and Jason Li, "Disinformation, Cybersecurity, and Energy Challenges," Stimson Center, September 12, 2019, https://www.stimson.org/2019/disinformation-cybersecurity-and-energy-challenges/.

[115] Emily Feng, "Taiwan Gets Tough On Disinformation Suspected From China Ahead Of Elections," *NPR*, December 6, 2019, https://www.npr.org/2019/12/06/784191852/taiwan-gets-tough-on-disinformation-suspected-from-china-ahead-of-elections; and Sam Byford, "Taiwan plans to ban major Chinese video streaming services," The Verge, August 20, 2020, https://www.theverge.com/2020/8/20/21376931/taiwan-ban-china-streaming-tencent-baidu-iqiyi-wetv.

[116] g0v website, accessed December 3, 2020, https://g0v.asia/; Andrew Leonard, "How Taiwan's Unlikely Digital Minister Hacked the Pandemic," *Wired*, July 23, 2020, https://www.wired.com/story/how-taiwans-unlikely-digital-minister-hacked-the-pandemic/; and Steven Butler and Iris Hsu, "Q&A: Taiwan's digital minister on combatting disinformation without censorship," Committee to Protect Journalists, May 23, 2019, https://cpj.org/2019/05/qa-taiwans-digital-minister-on-combatting-disinfor/.

[117] Jacob Changama and Jonas Parello-Plesner, "Taiwan's Disinformation Solution," *The American Interest*, February 6, 2020, https://www.the-american-interest.com/2020/02/06/taiwans-disinformation-solution/; DoubleThink Lab website, accessed December 18, 2020, https://doublethinklab.org/; Open Culture Foundation website, accessed December 18, 2020, https://ocf.tw/en/; and CoFacts website, accessed December 18, 2020, https://cofacts.g0v.tw/.

[118] Azeem Azhar, "How Taiwan is Using Technology to Foster Democracy (with Digital Minister Audrey Tang)," *Harvard Business Review*, October 14, 2020, https://hbr.org/podcast/2020/10/how-taiwan-is-using-technology-to-foster-democracy-with-digital-minister-audrey-tang; and Nicola Smith, "Schoolkids in Taiwan Will Now Be Taught How to Identify Fake News, *TIME Magazine*, April 7, 2017, https://time.com/4730440/taiwan-fake-news-education/.

[119] Y-Ting Lien, "Why China's COVID-19 Disinformation Campaign Isn't Working in Taiwan," *The Diplomat*, March 20, 2020, https://thediplomat.com/2020/03/why-chinas-covid-19-disinformation-campaign-isnt-working-in-taiwan/.

[120] Audrey Tang, "View from Taipei: Battling Disinformation and Hacking a Pandemic," event, Aspen Institute, August 7, 2020, https://www.youtube.com/watch?v=RLgJr40c08Q&feature=emb_logo&ab_channel=TheAspenInstitute.

Social media platforms and big tech responses

As various actors – foreign state-sponsored and others – have exploited social media platforms to push disinformation and influence campaigns, the platforms have not always acted as expeditiously as required to reduce the potential for harm. To be sure, addressing disinformation in its various manifestations is a complicated task and arguably one that the private sector should not take on unilaterally. However, big tech's integral role in facilitating the global information landscape also makes it the first line of defense against the manipulation of such information. Free speech considerations and structural financial incentives, among other things, have typically slowed social media platform responses to disinformation, but the COVID-19 pandemic and subsequent "infodemic" showed just how quickly the platforms could move when they deem a threat serious enough. Companies such as Facebook, Twitter, and Google took significantly more proactive steps than in past years to safeguard their platforms against misuse as COVID-19 spread, and the 2020 U.S. presidential election neared. While these responses continue to evolve, it's worth tracking key policies, strategies, and techniques implemented, some with direct implications for foreign state-sponsored disinformation operations.

The COVID-19 pandemic and subsequent "infodemic" **showed just how quickly the platforms could move when they deem a threat serious enough.**

Before the emergence of COVID-19 and the urgency it created, social media platforms had been confronted with election-related misinformation and disinformation across the globe, prompting revisions to their policies on content moderation or the creation of new ones altogether.[121] Increased scrutiny from policymakers, civil society groups, academics, and others have put pressure on social media platforms to shore up their election-related policies.[122] Twitter, for example, introduced a Civic Integrity Policy and Microsoft's Defending Democracy Program engages with diverse stakeholders globally to ensure the integrity of democratic processes, including defending against disinformation campaigns.[123]

Notably, several major platforms developed stronger labeling and removal policies targeted at state-controlled media or state-sponsored information operations, particularly concerning the 2020 U.S. election. In June 2020, Facebook and Instagram started labeling state-controlled media outlets and later blocked ads from those entities that targeted people in the United States.[124] Twitter took similar steps in August 2020 when it began labeling state-controlled and affiliated media accounts, representing one of the most comprehensive approaches toward state-entity labeling, as these labels appeared not only on entity accounts but also on actual tweets and within search results.[125] Labeling state-affiliated entities can provide users with additional context that could potentially

[121] Steven Overly, "What global elections have taught Silicon Valley about misinformation," *POLITICO*, October 23, 2020, https://www.politico.com/news/2020/10/23/what-global-elections-have-taught-silicon-valley-about-misinformation-431592; and Jon Lloyd, "Mozilla Sheds Light on Platform Election Policies," Mozilla, October 16, 2020, https://foundation.mozilla.org/en/blog/mozilla-sheds-light-platform-election-policies/.

[122] "Evaluating Platform Election-Related Speech Policies," Election Integrity Partnership, October 28, 2020, https://www.eipartnership.net/policy-analysis/platform-policies.

[123] "Civic integrity policy," Twitter, January 2021, https://help.twitter.com/en/rules-and-policies/election-integrity-policy; and "Defending Democracy Program," Microsoft, accessed January 9, 2021, https://news.microsoft.com/on-the-issues/topic/defending-democracy-program/.

[124] Nathaniel Gleicher, "Labeling State-Controlled Media On Facebook," Facebook, June 4, 2020, https://about.fb.com/news/2020/06/labeling-state-controlled-media/.

[125] "New labels for government and state-affiliated media accounts," Twitter, August 6, 2020, https://blog.twitter.com/en_us/topics/product/2020/new-labels-for-government-and-state-affiliated-media-accounts.html.

influence their engagement with the content.[126] However, these labels are not fully or consistently implemented across platforms, and the labeling criteria remains unclear.[127]

During 2020, platforms removed networks associated with foreign state-sponsored actors that pushed disinformation and influence operations.[128] In March 2020, Facebook began publishing monthly reports related to its policy on Coordinated Inauthentic Behavior, in which it discloses coordinated inauthentic networks discovered and removed from both Facebook and Instagram.[129] Twitter has also removed accounts it determines are state-backed information operations and makes them publicly available in its archives.[130] Since 2018, Google's Threat Analysis Group has been working to detect state-sponsored phishing and hacking attempts and identify influence operations by foreign governments, such as Russia, China, and Iran.[131] For example, thousands of YouTube channels were terminated in 2020 as part of the Threat Analysis Group's ongoing investigation into coordinated influence operations linked to China.[132]

The COVID-19 pandemic arguably precipitated a shift in platform practices in removing content, especially conspiracy theories, hate speech, coordinated information campaigns, and other misleading information that could lead to harm.[133] Platforms issued a joint industry statement in March 2020 pledging to combat COVID-19 misinformation and later pledged to remove false information related to COVID-19 vaccines, highlighting the industry's public commitment in addressing mis- and disinformation.[134]

Platforms expanded and further invested in previously established third-party fact-checking partnerships and programs and worked toward minimizing exposure to misinformation and providing verified and educational information, such as a Twitter search prompt for coronavirus information and an "Election Hub."[135] Facebook pledged more than US$100 million to support journalists and fact-checkers and now

[126] Todd C. Helmus, James V. Marrone, Marek N. Posard, and Danielle Schlang, "Russian Propaganda Hits Its Mark: Experimentally Testing the Impact of Russian Propaganda and Counter-Interventions," RAND, 2020, https://www.rand.org/pubs/research_reports/RRA704-3.html.

[127] Nicole Buckley, Morgan Wack, Joey Schafer, and Martin Zhang, "Inconsistencies in State-Controlled Media Labeling," Election Integrity Partnership, October 6, 2020, https://www.eipartnership.net/policy-analysis/inconsistencies-in-state-controlled-media-labeling.

[128] Jack Stubbs and Christopher Bing, "Facebook, Twitter dismantle global array of disinformation networks," *Reuters*, October 8, 2020, https://www.reuters.com/article/cyber-disinformation-facebook-twitter/facebook-twitter-dismantle-global-array-of-disinformation-networks-idINKBN26T2XF.

[129] "Coordinated Inauthentic Behavior," Facebook, December 6, 2018, https://about.fb.com/news/tag/coordinated-inauthentic-behavior/.

[130] "Information Operations," Twitter, accessed January 9, 2021, https://transparency.twitter.com/en/reports/information-operations.html.

[131] Kent Walker, "An update on state-sponsored activity," Google, August 23, 2018, https://blog.google/technology/safety-security/update-state-sponsored-activity/.

[132] "Threat Analysis Group," Google, last updated March 31, 2021, https://blog.google/threat-analysis-group/.

[133] Spandana Singh and Koustubh "K.J." Bagchi, "How Internet Platforms Are Combating Disinformation and Misinformation in the Age of COVID-19," New America, June 1, 2020, https://www.newamerica.org/oti/reports/how-internet-platforms-are-combating-disinformation-and-misinformation-age-covid-19/.

[134] Catherine Shu and Jonathan Shieber, "Facebook, Reddit, Google, LinkedIn, Microsoft, Twitter and YouTube issue joint statement on misinformation," *Tech Crunch*, March 16, 2020, https://techcrunch.com/2020/03/16/facebook-reddit-google-linkedin-microsoft-twitter-and-youtube-issue-joint-statement-on-misinformation/; and Rebecca Heilweil, "Twitter joins Facebook and YouTube in banning Covid-19 vaccine misinformation," Vox, December 16, 2020, https://www.vox.com/recode/22179145/twitter-misinformation-covid-19-vaccines-pfizer-moderna.

[135] Spandana Singh and Koustubh "K.J." Bagchi, "How Internet Platforms Are Combating Disinformation and Misinformation in the Age of COVID-19" New America, June 1, 2020, https://www.newamerica.org/oti/reports/how-internet-platforms-are-combating-disinformation-and-misinformation-age-covid-19/; "Coronavirus: Staying safe and informed on Twitter," Twitter, January 12, 2021, https://blog.twitter.com/en_us/topics/company/2020/covid-19.html; and Andrew Hutchinson, "Twitter Launches 2020 US Election Hub to Distribute Accurate, Timely Voter Information," *Social Media Today*, September 15, 2020, https://www.socialmediatoday.com/news/twitter-launches-2020-us-election-hub-to-distribute-accurate-timely-voter/585272/.

highlights authoritative sources on important topics, including COVID-19, the Holocaust, and elections. It has also added more educational resources to Instagram search and chatbots on WhatsApp to directly provide users with authoritative information.[136] In addition to providing increased funding for fact-checkers, Google launched an experimental platform to help journalists and fact-checkers quickly verify images and spot doctored images.[137] YouTube has fact-checking information panels from independent third-party publishers currently available in select countries, including Brazil, India, and the United States.[138]

Platforms also implemented new features and techniques to limit the spread of disinformation through content moderation and removal. Facebook and Twitter implemented significantly more content interventions in recent years than other platforms, removing more than 317,000 accounts and pages from January 2019 to November 2020.[139] However, COVID-19 restrictions forced many platforms to rely more heavily on machine moderation than ever before, further exposing blind spots in disparate enforcement depending on the nature and origin of the content.[140] In some cases, harmful material was more likely to slip past the machines while authoritative content was accidentally removed, and in other situations, incorrect takedowns occurred more often.[141] Moreover, automated moderation algorithms could be counterproductive, such as Twitter's adding fact-checking labels to all tweets mentioning 5G and COVID-19, which could enable the spread of conspiracy theories about those topics.[142]

[136] Campbell Brown, "Facebook Invests Additional $100 Million To Support Journalism Industry During Coronavirus Pandemic," Facebook Journalism Project, March 30, 2020, https://www.facebook.com/journalismproject/coronavirus-update-news-industry-support; Guy Rosen, "An Update on Our Work to Keep People Informed and Limit Misinformation About COVID-19," Facebook, April 16, 2020, https://about.fb.com/news/2020/04/covid-19-misinfo-update/; Matt O'Brien, "Facebook bans Holocaust denial, distortion posts," *AP*, October 12, 2020, https://apnews.com/article/election-2020-media-social-media-elections-mark-zuckerberg-14e8073ce6f7bd2a674c99ac7bbfc240; Guy Rosen, "Preparing for Election Day," Facebook, October 7, 2020, https://about.fb.com/news/2020/10/preparing-for-election-day/; "Keeping People Informed, Safe, and Supported on Instagram," Instagram, March 24, 2020, https://about.instagram.com/blog/announcements/coronavirus-keeping-people-safe-informed-and-supported-on-instagram; and Ivan Mehta, "World Health Organization's WhatsApp bot texts you coronavirus facts," *The Next Web*, March 20, 2020, https://thenextweb.com/apps/2020/03/20/world-health-organizations-whatsapp-bot-texts-you-coronavirus-facts/.

[137] Alexios Mantzarlis, "COVID-19: $6.5 million to help fight coronavirus misinformation," Google, April 2, 2020, https://blog.google/outreach-initiatives/google-news-initiative/covid-19-65-million-help-fight-coronavirus-misinformation/; and Karen Hao, "Google has released a tool to spot faked and doctored images," *MIT Technology Review*, February 5, 2021, https://www.technologyreview.com/2020/02/05/349126/google-ai-deepfakes-manipulated-images-jigsaw-assembler.

[138] "See fact checks in YouTube search results," Google, accessed on January 12, 2021, https://support.google.com/youtube/answer/9229632?hl=en-GB&ref_topic=9257092.

[139] Kamya Yadav, "Platform Interventions: How Social Media Counters Influence Operations," Carnegie Endowment for International Peace, January 25, 2021, https://carnegieendowment.org/2021/01/25/platform-interventions-how-social-media-counters-influence-operations-pub-83698; and Samantha Bradshaw, Hannah Bailey, and Philip N. Howard, "Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation," Oxford Internet Institute, January 13, 2021, https://comprop.oii.ox.ac.uk/research/posts/industrialized-disinformation/.

[140] Marc Faddoul, "COVID-19 is triggering a massive experiment in algorithmic content moderation," Brookings Institution, April 28, 2020, https://www.brookings.edu/techstream/covid-19-is-triggering-a-massive-experiment-in-algorithmic-content-moderation/.

[141] Mark Scott and Laura Kayali, "What happened when humans stopped managing social media content," *POLITICO*, October 21, 2020, https://www.politico.eu/article/facebook-content-moderation-automation/; and James Vincent, "YouTube brings back more human moderators after AI systems over-censor," *The Verge*, September 21, 2020, https://www.theverge.com/2020/9/21/21448916/youtube-automated-moderation-ai-machine-learning-increased-errors-takedowns.

[142] Paige Leskin, "Twitter has apologized for slapping a COVID-19 label on tweets about 5G, but experts say the platform's algorithm could be encouraging the spread of conspiracy theories," *Business Insider*, June 29, 2020, https://www.businessinsider.com/twitter-5g-coronavirus-label-blames-algorithm-encourages-conspiracy-theories-2020-6.

Following concerns over its lack of transparency regarding content moderation, TikTok developed content removal policies around misleading information and disinformation campaigns and joined other major platforms in signing the European Commission Code of Practice on Disinformation.[143] TikTok also increased its efforts to remove videos with demonstrably false content while displaying warnings on videos with questionable content, claiming that these labels decreased the rate at which users share unsubstantiated content by 24 percent and likes by 7 percent.[144]

Changes to content recommendation algorithms featured heavily in platform responses in 2020.

Meanwhile, Reddit has been incorporating a hybrid and decentralized approach. It employs a small centralized team of moderators, while users known as "Mods" voluntarily moderate the majority of content in individual subreddits and have significant editorial discretion to remove content and mute or ban specific users.[145] In April 2020, Reddit added an AutoModerator tool, a built-in bot providing basic algorithmic tools that allow users to identify and remove obvious forms of misinformation in subreddits, which they can also report to the platform.[146]

Changes to content recommendation algorithms featured heavily in platform responses in 2020, though platforms continue to face criticism over such algorithms, which have been found to promote disinformation and harmful content.[147] Platforms sought to address this problem by making changes to their algorithms to minimize user exposure to potentially misleading and false information. As one of the biggest culprits, YouTube now reduces recommendations for what it categorizes as "borderline content," content that could potentially misinform users.[148] While these changes reduced the number of conspiracy theory videos recommended, they have not fully limited exposure to misinformation and disinformation.[149] Facebook made its content recommendation guidelines public in August 2020, after coming under fire for the role

[143] Hannah Murphy and Yuan Yang, "TikTok rushes to build moderation teams as concerns rise over content," *Irish Times*, December 20, 2019, https://www.irishtimes.com/business/technology/tiktok-rushes-to-build-moderation-teams-as-concerns-rise-over-content-1.4121460; Lavanya Mahendran and Nasser Alsherif, "Adding clarity to our Community Guidelines," TikTok, January 8, 2020, https://newsroom.tiktok.com/en-us/adding-clarity-to-our-community-guidelines; and "TikTok Signs Up to EU Initiative to Fight Disinformation," Dot Europe, June 22, 2020, https://doteurope.eu/news/tiktok-signs-up-to-eu-initiative-to-fight-disinformation/.

[144] Gina Hernandez, "New prompts to help people consider before they share," TikTok, February 3, 2021, https://newsroom.tiktok.com/en-us/new-prompts-to-help-people-consider-before-they-share.

[145] Spandana Singh, "Everything in Moderation: An Analysis of How Internet Platforms Are Using Artificial Intelligence to Moderate User-Generated Content," New America, July 22, 2019, https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/.

[146] Singh, "Everything in Moderation."

[147] Katherine J. Wu, "Radical ideas spread through social media. Are the algorithms to blame?" *PBS Nova*, March 28, 2019, https://www.pbs.org/wgbh/nova/article/radical-ideas-social-media-algorithms/.

[148] Craig Timberg, Drew Harwell, and Tony Romm, "YouTube excels at recommending videos – but not at detecting hoaxes," *Washington Post*, February 22, 2018, https://www.washingtonpost.com/business/technology/youtube-excels-at-recommending-videos—but-not-at-detecting-hoaxes/2018/02/22/6063268e-1803-11e8-92c9-376b4fe57ff7_story.html; and "Continuing our work to improve recommendations on YouTube," YouTube, January 25, 2019, https://blog.youtube/news-and-events/continuing-our-work-to-improve.

[149] Charlotte Jee, "YouTube has nearly halved the number of conspiracy theory videos it recommends," *MIT Technology Review*, March 3, 2020, https://www.technologyreview.com/2020/03/03/905565/youtube-halved-conspiracy-theory-videos-recommends/; Hanaa Tameez, "YouTube's algorithm is pushing climate misinformation videos, and their creators are profiting from it," NiemanLab, January 16, 2020, https://www.niemanlab.org/2020/01/youtubes-algorithm-is-pushing-climate-misinformation-videos-and-their-creators-are-profiting-from-it/; and Greg Bensinger, " YouTube says viewers are spending less time watching conspiracy theory videos. But many still do," *Washington Post*, December 3, 2019, https://www.washingtonpost.com/technology/2019/12/03/youtube-says-viewers-are-spending-less-time-watching-conspiracy-videos-many-still-do/.

its algorithm played in fostering conspiracy theories on the platform, listing specific content categories that are not eligible for recommendations, including false or misleading content.[150]

As platforms continue to implement policies and strategies to counter false and manipulated content, disinformation actors likewise adapt tactics to evade detection and increase their efficacy. These tactics include the use of encrypted messaging apps and microtargeting to personalize disinformation narratives and blur the lines between content that was organically created and spread versus content created and circulated by external agitators.

As new topics arise and information dissemination evolves, **it is imperative to understand which platform's counter-disinformation responses show the most promise and how they might scale.**

Rising levels of offline violence in countries such as Brazil and India have been attributed to the spread of disinformation on messaging applications such as WhatsApp and Telegram.[151] The encrypted nature of these platforms makes it hard to identify and counteract the spread of false information and to understand the full extent of the problem.[152] Following a series of violent incidents prompted by rumors spread on its platform, WhatsApp reduced the number of chats to which a user can forward a message simultaneously.[153] WhatsApp claimed this intervention resulted in a 70 percent reduction globally in the number of highly forwarded messages. In September 2020, Facebook Messenger also began implementing similar forwarding limits.[154] Other encrypted messaging apps, such as Telegram, face similar challenges and could potentially use WhatsApp as a model for limiting the spread of misinformation and disinformation on their platforms.[155]

Finally, platforms are improving their ability to detect synthetic and manipulated media, such as deepfakes, which provide disinformation actors new opportunities to spread confusion and distrust online.[156] In February 2020, Twitter introduced new rules and labels for synthetic and manipulated media requiring the removal of manipulated content that is deceptive by nature.[157] In September 2020, Google's emerging threats unit Jigsaw released a large dataset of visual deepfakes

[150] Sarah Perez, "Facebook partially documents its content recommendation system," *TechCrunch*, August 31, 2020, https://techcrunch.com/2020/08/31/facebook-partially-documents-its-content-recommendation-system/.

[151] Samuel Woolley, "Encrypted messaging apps are the future of propaganda," Brookings Institution, May 1, 2020, https://www.brookings.edu/techstream/encrypted-messaging-apps-are-the-future-of-propaganda/.

[152] Kamya Yadav, "Platform Interventions: How Social Media Counters Influence Operations," Carnegie Endowment for International Peace, January 25, 2021, https://carnegieendowment.org/2021/01/25/platform-interventions-how-social-media-counters-influence-operations-pub-83698.

[153] Timothy McLaughlin, "How WhatsApp Fuels Fake News and Violence in India," *Wired*, December 12, 2018, https://www.wired.com/story/how-whatsapp-fuels-fake-news-and-violence-in-india/; and Alex Hern," WhatsApp to restrict message forwarding after India mob lynchings," *The Guardian*, July 20, 2018, https://www.theguardian.com/technology/2018/jul/20/whatsapp-to-limit-message-forwarding-after-india-mob-lynchings.

[154] Jay Sullivan, "Messenger Launches Forwarding Limits," Facebook Messenger, September 3, 2020, https://messengernews.fb.com/2020/09/03/messenger-launches-forwarding-limits/.

[155] Alexandra S. Levine, "Telegram surfaces as preferred app of extremist rioters," *POLITICO*, June 4, 2020, https://www.politico.com/newsletters/morning-tech/2020/06/04/telegram-surfaces-as-preferred-app-of-extremist-rioters-788230.

[156] Alex Engler, "Fighting deepfakes when detection fails," Brookings Institution, November 14, 2020, https://www.brookings.edu/research/fighting-deepfakes-when-detection-fails/; and Tim Hwang, "Deepfakes: A Grounded Threat Assessment," Center for Security and Emerging Technology, July 2020, https://cset.georgetown.edu/research/deepfakes-a-grounded-threat-assessment/.

[157] Yoel Roth and Ashita Achuthan, "Building rules in public: Our approach to synthetic & manipulated media," Twitter, February 4, 2020, https://blog.twitter.com/en_us/topics/company/2020/new-approach-to-synthetic-and-manipulated-media.html.

available gratis to the research community for use in developing synthetic video detection methods.[158] As synthetic media quality improves, investing in the ability to spot and counteract such media through research is a key counter-disinformation strategy. YouTube regularly updates its deceptive practice policies, which include removing content that has been manipulated or doctored to mislead users and terminating channels that artificially increase viewership metrics.[159] In August 2020, TikTok expanded its recently formulated community guidelines to include a policy that prohibits deceptive or misleading synthetic or manipulated content.[160]

As new topics arise and information dissemination evolves, it is imperative to understand which platform's counter-disinformation responses show the most promise and how they might scale.

This requires platforms to work with industry experts, researchers, and civil society to ensure a transparent assessment of platform misuse and the efficacy of responses to date. The next section explores what civil society actors are doing on this front.

CASE STUDY: PINTEREST

Facing a growing movement of anti-vaccination entrepreneurs, Pinterest adopted a zero-tolerance strategy on misinformation, implementing some of the strongest and most proactive measures of any platform.[161] Its health misinformation policy states that any content that could result in immediate, adverse effects on someone's health or the general public's safety has no place on the platform.[162] There are no exceptions for prominent political leaders or celebrities. Its content moderation team also used AI to run proactive searches around COVID-19-related misinformation and removed these pins.[163] Despite these efforts, some health misinformation remains on the platform, highlighting the difficulty of addressing this issue.[164]

[158] Nick Dufour and Andrew Gully, "Contributing Data to Deepfake Detection Research," Google AI Blog, September 24, 2019, https://ai.googleblog.com/2019/09/contributing-data-to-deepfake-detection.html.

[159] "Spam, deceptive practices, & scams policies," YouTube Help, accessed January 12, 2021, https://support.google.com/youtube/answer/2801973?hl=en.

[160] Vanessa Pappas, "Combating misinformation and election interference on TikTok," TikTok, August 5, 2020, https://newsroom.tiktok.com/en-us/combating-misinformation-and-election-interference-on-tiktok.

[161] Erin Brodwin, "How Pinterest beat back vaccine misinformation — and what Facebook could learn from its approach," *State News*, September 21, 2020, https://www.statnews.com/2020/09/21/pinterest-facebook-vaccine-misinformation/.

[162] "Health misinformation," Pinterest, accessed on January 12, 2021, https://help.pinterest.com/en/article/health-misinformation.

[163] Kyle Wiggers, "Pinterest launches Today tab with curated topics, fights coronavirus misinformation with AI," *VentureBeat*, March 24, 2020, https://venturebeat.com/2020/03/24/pinterest-launches-today-tab-with-curated-topics-fights-coronavirus-misinformation-with-ai/.

[164] Harrison Weber, "Pinterest continues its crackdown on health misinformation," *Fast Company*, August 28, 2019, https://www.fastcompany.com/90396831/pinterest-continues-cracking-down-on-anti-vaxxer-misinformation.

In September 2020, Pinterest announced new policies on election-related misinformation, reinforcing its community guidelines to prohibit false or misleading content that impedes an election's integrity or an individual or group's civic participation.[165] Moreover, in addition to its 2018 ban on political ads, Pinterest's new policies limit recommendations about election-related content (like election memes or slogans) in places like the home feed and turn off search autocomplete and search guides for specific election-related terms.[166]

CASE STUDY: WIKIPEDIA

Wikipedia accounts for 55 million articles in more than 300 languages, with a monthly global reach of 1.5 billion.[167] Its content is almost entirely moderated by more than 280,000 volunteer editors worldwide who crowdsource efforts to authenticate information.[168] The crowdsourced nature of Wikipedia's editing, fact-checking, and content moderation contributes to a model that provides access to free and open information while combating false information on the platform.[169] For known controversial topics, such as politics and religion, dedicated editors track edits in real time to counter potential disinformation. Overall, this networked system of review helps to produce consensus and promote authoritative content. In addition to automated moderation, the platform uses community-driven fact-checking and requires sources and citations for all factual statements.[170] Wikipedia's approach contributes to a high level of transparency and has resulted in the platform largely staying above the misinformation fray.[171]

During the pandemic, this model helped Wikipedia stay abreast of disinformation, as it could quickly verify information with its network of public health experts.[172] These editors had already

[165] "Our Commitment to Election Integrity and Civic Engagement," *NewsRoom*, September 23, 2020, https://newsroom.pinterest.com/en/electionsintegrity; and "Community guidelines," Pinterest, accessed January 12, 2021, https://policy.pinterest.com/en/community-guidelines.

[166] I. Bonifacic, "Pinterest users won't see ads when they search for election-related content," *Engadget*, September 3, 2020, https://www.engadget.com/pinterest-election-2020-policies-212046383.html.

[167] "Wikipedia," Wikipedia, accessed January 10, 2021, https://en.wikipedia.org/wiki/Wikipedia.

[168] Katherine Maher and Manoush Zomorodi, "NYC Media Lab Celebrates Wikipedia's 20th Birthday (full video)," event, NYC Media Lab, January 15, 2021, https://www.youtube.com/watch?v=qyJ9j8yesis&ab_channel=nycmedialab.

[169] Zachary J. McDowell and Matthew A. Vetter, "It Takes a Village to Combat a Fake News Army: Wikipedia's Community and Policies for Information Literacy," *Social Media + Society*, July 2020, https://journals.sagepub.com/doi/pdf/10.1177/2056305120937309.

[170] "Wikipedia: Automated moderation," Wikipedia, accessed January 4, 2021, https://en.wikipedia.org/wiki/Wikipedia:Automated_moderation; and Alex Pasternak, "How Wikipedia's volunteers became the web's best weapon against misinformation," *Fast Company*, March 7, 2020, https://www.fastcompany.com/90471667/how-wikipedia-volunteers-became-the-webs-best-weapon-against-misinformation.

[171] Pasternak, "How Wikipedia's volunteers became the web's best weapon against misinformation."

[172] WikiProject Medicine," Wikipedia, accessed January 4, 2021, https://en.wikipedia.org/wiki/Wikipedia:WikiProject_Medicine.

been stress tested during the 2014 Ebola crisis and had therefore gained the necessary skills to cover a rapidly changing information environment, ensure high quality across multiple languages, and respond to emerging information vacuums. During initial coverage of the virus, editors mapped relevant articles, and bots tracked edits in real time. A dedicated placeholder was created for the COVID-19 vaccine before its development to counter emerging conspiracies, showing an innovative approach to respond to information vacuums.

Wikipedia's community and source-driven approach has led to other platforms like Facebook using it as an authoritative source, validating its success at countering false information.[173] Wikipedia complements traditional media in its ability to provide a holistic perspective to a given topic. Moreover, its archive of articles and past edits is key to uncovering how people's understanding of certain topics evolves over time, as well as the interconnectedness and global distribution of those topics.

Civil Society Responses

Civil Society Organizations (CSOs) are a key player in helping mitigate vulnerabilities and strengthen resilience against disinformation. Indeed, a December 2020 global report from the Carnegie Endowment for International Peace found that nearly half of the initiatives to counter influence operations emanate from civil society; the National Endowment for Democracy mapped out 175 CSOs working in the counter-disinformation space worldwide in 2021.[174]

These organizations often bring localized experiences and lessons learned to advocate and apply pressure to governments, social media and tech companies, businesses, and others to tackle disinformation in a way that upholds freedom of expression and other fundamental rights.[175] In some cases, they're best placed to work with governments to improve trust in institutions, which can help mitigate the polarization that facilitates the creation and sustains the resonance of disinformation. Given that the best defense against digital disinformation is to address the policy and societal issues exploited by disinformation operations, civil society actors are also well-placed to help address those issues.

Civil Society Organizations conduct critical research and training and provide valuable recommendations to governments on how to effectively counter disinformation. One such initiative is the EU DisinfoLab, which tackles disinformation campaigns targeting the EU, its member states, core institutions, and core values. Its policy and research efforts have been commissioned by the European Parliament and the European

[173] Omer Benjakob, "There's a lot Wikipedia can teach us about fighting disinformation," *Wired*, June 8, 2019, https://www.wired.co.uk/article/wikipedia-fake-news-disinformation.

[174] Victoria Smith, "Mapping Worldwide Initiatives to Counter Influence Operations," Carnegie Endowment for International Peace, December 14, 2020, https://carnegieendowment.org/2020/12/14/mapping-worldwide-initiatives-to-counter-influence-operations-pub-83435#initiatives/; and Samantha Bradshaw and Lisa-Maria Neudert, "The Road Ahead: Mapping Civil Society Responses to Disinformation," National Endowment for Democracy, January 25, 2021, https://www.ned.org/mapping-civil-society-responses-to-disinformation-international-forum.

[175] "Freedom and Accountability: A Transatlantic Framework for Moderating Speech Online (Philadelphia: Annenberg Public Policy Center, June 2020), https://www.annenbergpublicpolicycenter.org/feature/transatlantic-working-group-freedom-and-accountability/.

Commission, among others.[176] Using open-source intelligence and social media network analysis, the organization identifies, uncovers, and explains disinformation campaigns and networks, such as a long-standing influence operation ("Indian Chronicles") targeting international institutions like the UN and EU to promote Indian interests.[177]

Acting as watchdogs, CSOs monitor social media, exposing and refuting disinformation as it appears.[178] Their work involves proactive measures to raise awareness of the tactics and techniques used to create and propagate disinformation, known as "prebunking," as well as reactive measures to analyze, verify, and, when necessary, debunk specific narratives. The International Fact-Checking Network (IFCN), for example, brings together fact-checking organizations to ensure the application of basic standards through a signed Code of Principles.[179] In 2016, IFCN partnered with Facebook's Third-Party Fact-Checking network to verify stories shared on users' newsfeeds.[180] As a result of the COVID-19 infodemic, IFCN launched the #CoronaVirusFacts Alliance, which unites more than 100 fact-checkers worldwide in publishing, sharing, and translating accurate information about the virus.[181] Similarly, in March 2020, Argentina's *Chequeado* coordinated with other regional fact-checking organizations to establish the Latam *Chequea* Coronavirus project, which unites more than 25 fact-checking organizations in the region to address the infodemic.[182]

Additionally, CSOs help to raise general awareness of disinformation and amplify media literacy programs that educate citizens so they have the tools to protect themselves. They work with local and independent media to educate them on protecting their reporting and tradecraft from malign foreign influence. This includes pushing for standards in how the journalistic community responds to leaks of hacked information and sharing verification good practices. First Draft News, for example, has been leading efforts to build news media's capacity to verify and authenticate media content globally. Through its CrossCheck model, the organization pioneered collaborative reporting around elections in the United States, France, the United Kingdom, Germany, Brazil, Nigeria, Spain, and the EU, showing that competing newsrooms can work together for more effective, efficient, and responsible reporting.[183] According to research, audiences reported higher levels of trust in the reporting due to multiple newsrooms collaborating on stories and felt that CrossCheck was more independent, impartial, and credible because it included so many outlets.[184]

[176] "About Us," EU DisinfoLab website, accessed January 4, 2021, https://www.disinfo.eu/about-us; and European Parliamentary Research Service, "Automated tackling of disinformation" (Brussels: EPRS, March 2019).

[177] Gary Machado, Alexandre Alaphilippe, and Roman Adamczyk, "Indian Chronicles: Deep Dive Into a 15-year Operation Targeting the EU and UN to Serve Indian Interests," EU DisinfoLab, December 9, 2020, https://www.disinfo.eu/publications/indian-chronicles-deep-dive-into-a-15-year-operation-targeting-the-eu-and-un-to-serve-indian-interests.

[178] Bret Schafer, "A Democratic Response to Digital Disinformation: The Role of Civil Society," Johns Hopkins University American Institute for Contemporary German Studies, September 28, 2018, https://www.aicgs.org/2018/09/a-democratic-response-to-digital-disinformation-the-role-of-civil-society/.

[179] "The International Fact-Checking Network," Poynter Institute, accessed January 10, 2021, https://www.poynter.org/ifcn/; and "Commit to transparency — sign up for the International Fact-Checking Network's code of principles," Poynter Institute, IFCN Code of Principles, accessed January 10, 2021, https://ifcncodeofprinciples.poynter.org/.

[180] "Facebook's Third-Party Fact-Checking Program," Facebook Journalism Project, accessed February 10, 2021, https://www.facebook.com/journalismproject/programs/third-party-fact-checking.

[181] "Fighting the Infodemic: The #CoronaVirusFacts Alliance," Poynter Institute, February 10, 2021, https://www.poynter.org/coronavirusfactsalliance/.

[182] "Información chequeada sobre el Coronavirus," Latam Chequea, accessed February 10, 2021, https://chequeado.com/latamcoronavirus/.

[183] "Spotlight - First Draft," Disinfo Cloud, May 15, 2020, https://disinfocloud.com/blog/spotlight-firstdraft.

[184] "Research on CrossCheck journalists and readers suggests positive impact for project," *First Draft*, November 16, 2017, https://firstdraftnews.org/latest/crosscheck-qualitative-research.

Similarly, Jordan-based Fatabyyno monitors and debunks disinformation in 18 countries across the Middle East and North Africa. The Iraqi non-governmental organization (NGO) Tech 4 Peace conducts training across Iraq to teach citizens and journalists how to verify news sources; during protests in 2019, they worked to debunk false information amid internet restrictions and social media bans imposed by the government.[185]

Civil society actors are also playing a role in detecting online manipulation and disinformation operations. In Lithuania, *Demaskuok* searches the information ecosystem of the Baltic states for disinformation and propaganda, using an AI tool to flag suspected disinformation and a network of human fact-checkers to verify or discredit flagged content. The media-led initiative claims to have reached 90 percent of Lithuania's population.[186] In addition, the citizen-led "Baltic Elves" initiative has gained recognition in recent years for its army of volunteers dedicated to tracing trolls and challenging Russian propaganda online.[187] Inspired by Lithuania's Baltic Elves, a group referred to as the Czech Elves is combating disinformation and propaganda ranging from tracking down the originators of disinformation online to exposing trolls on social media networks and mapping chain emails.[188] Additionally, UK-based Bellingcat comprises a distributed team of investigative journalists, researchers, and citizens specializing in open-source intelligence. Their online investigation toolkit compiles hundreds of useful online tools for open-source intelligence in one place, so independent researchers and journalists globally are empowered to perform investigations and digital forensic analysis.[189]

Despite these critical efforts, civil society approaches are often limited by a lack of sustainable funding and resources that hinder their ability to proactively detect and confront disinformation operations and actors.[190] Nevertheless, there are a growing number of initiatives to help CSOs hone their disinformation detection capabilities. For example, the Technology and Social Change project created the Media Manipulation Casebook to serve as a research platform that advances knowledge of misinformation and disinformation and their threats to democracy, public health, and security.[191] The casebook provides strategies and case studies for civil society and other groups to counter misinformation and disinformation.[192] Other resources include a free online book that draws on the knowledge and

[185] Mark Stencel and Joel Luther, "From Toronto to New Delhi, fact-checkers find reinforcements," Duke Reporters Lab, September 16, 2019, https://reporterslab.org/tag/fatabyyano/; Fatabyyno website, accessed December 16, 2020, https://fatabyyano.net; and "Tech-savvy activists debunk fake news engulfing Iraq protests," *France24*, October 26, 2019, https://www.france24.com/en/20191026-tech-savvy-activists-debunk-fake-news-engulfing-iraq-protests.

[186] Benas Gerdziunas, "Lithuania hits back at Russian disinformation," *DW*, September 27, 2018, https://www.dw.com/en/lithuania-hits-back-at-russian-disinformation/a-45644080.

[187] Kim Sengupta, "Meet the Elves, Lithuania's digital citizen army confronting Russian trolls," *The Independent*, July 27, 2019, https://www.independent.co.uk/news/world/europe/lithuania-elves-russia-election-tampering-online-cyber-crime-hackers-kremlin-a9008931.html.

[188] Adam Zamecnik, "An army of volunteer 'elves' fights disinfo in the Czech Republic," *Coda*, May 19, 2020, https://www.codastory.com/disinformation/volunteers-fight-disinfo-czech-republic/.

[189] Bellingcat's Online Investigative Toolkit, accessed December 16, 2020, https://docs.google.com/spreadsheets/d/18rtqh8EG2q1xBo2cLNyhIDuK9jrPGwYr9DI2UncoqJQ/edit#gid=930747607.

[190] Carl Miller and Chloe Colliver, "Developing a Civil Society Response to Online Manipulation," Institute for Strategic Dialogue, August 13, 2020, https://www.isdglobal.org/isd-publications/developing-a-civil-society-response-to-online-manipulation/.

[191] "Technology and Social Change Research Project Team," Harvard Kennedy School Shorenstein Center on Media, Politics and Public Policy, accessed December 8, 2020, https://shorensteincenter.org/programs/technology-social-change/.

[192] Joan Donovan, "How Civil Society Can Combat Misinformation and Hate Speech Without Making It Worse," *Medium*, September 28, 2020, https://medium.com/political-pandemonium-2020/how-civil-society-can-combat-misinformation-and-hate-speech-without-making-it-worse-887a16b8b9b6.

experience of top journalists and researchers in the field to provide tools and techniques to execute investigations into digital media manipulation, disinformation, and information operations.[193] Similarly, the Institute for Strategic Dialogue's toolkit lays out an approach that organizations can use to track online disinformation.[194] The process is intended to have a low barrier to entry, with each stage achievable using either over-the-counter or free-to-use social listening tools.

CASE STUDY: COLLABORATIVE FACT-CHECKING

In the immediate aftermath of the 2017 Puebla earthquake in Mexico, an ad hoc, citizen-led initiative organized and corroborated information to strengthen the humanitarian response and provide verified information about the earthquake.[195] Composed of media outlets, companies, NGOs, and universities, the initiative led to the creation of the hashtag *#Verificado19S* and the Twitter account *@Verificado19S*. Inspired by the impact of *Verificado19S*, Animal Político, AJ+ Español, Pop-Up Newsroom, and others organized an initiative called *Verificado* 2018 to counter false information related to Mexico's 2018 elections. They brought together more than 60 fact-checking organizations, media outlets, and CSOs in 28 of Mexico's 32 states to respond to individual queries about the accuracy of election-related reports.[196]

Verificado 2018 represents a model for fighting disinformation in Mexico, having garnered 5.4 million visits to its 400 published entries and engaging voters through different channels, including its WhatsApp group, which had 9,600 subscriptions and 60,700 interactions by 2019.[197] In early 2020, the team that participated in *Verificado19S* started *Verificovid* to tackle false information surrounding the coronavirus in Mexico.[198] The team of journalists, communicators, designers, and doctors monitors and debunks false information and claims circulating on social networks and digital media and has already gained more than 65,000 followers on Twitter by the end of March 2021.

[193] "Verification Handbook For Disinformation And Media Manipulation," European Journalism Centre, accessed December 10, 2020, https://datajournalism.com/read/handbook/verification-3.

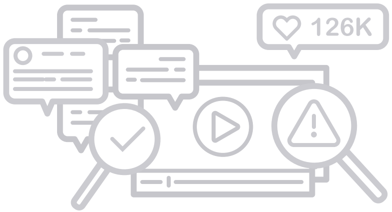
[194] Carl Miller and Chloe Colliver, "Disinformation Starter Kit: The 101 of Disinformation Detection," Institute for Strategic Dialogue, August 13, 2020, https://www.isdglobal.org/isd-publications/the-101-of-disinformation-detection/.

[195] *Verificado 19S*, accessed January 8, 2021, https://verificado19s.org/; and "How can misinformation be addressed during crises?," Disinfo Cloud, July 10, 2020, https://disinfocloud.com/blog/research-disaster-misinformation.

[196] *Verificado 18*, accessed January 8, 2021, https://verificado.mx/; and Lauren Hazard Owen, "WhatsApp is a black box for fake news. Verificado 2018 is making real progress fixing that," NiemanLab, June 1, 2018, https://www.niemanlab.org/2018/06/whatsapp-is-a-black-box-for-fake-news-verificado-2018-is-making-real-progress-fixing-that/.

[197] Luiza Bandeira, Donara Barojan, Roberta Braga, Jose Luis Peñarredonda, and Maria Fernanda Pérez Argüello, "Disinformation in Democracies: Strengthening Digital Resilience in Latin America," Atlantic Council, March 28, 2019, https://www.atlanticcouncil.org/in-depth-research-reports/report/disinformation-democracies-strengthening-digital-resilience-latin-america/.

[198] "#QuédateEnCasa," *Verificovid*, accessed January 8, 2021, https://verificovid.mx/.



4. What’s the State of Technology?

Purveyors of disinformation rely on a number of digital tools to produce and spread online disinformation.[199] These tools include manipulated or synthetic media such as deepfakes and less sophisticated “cheapfakes,” visual content such as memes, manufactured amplification to artificially boost content, and trolling by deliberately posting offensive or inflammatory content to provoke readers or disrupt conversations.[200]

The proliferation of online manipulation and the viral spread of inauthentic content, exacerbated by the COVID-19 pandemic, is an evolving challenge.[201] Despite efforts to curb the spread of bot-generated content and engagement, bots are becoming more human-like and difficult to detect.[202] Furthermore, as major social media platforms become inhospitable to purveyors of disinformation, malign actors increasingly use encrypted messaging apps and closed networks to spread disinformation and propaganda.[203] These networks will be particularly important as social media platforms like Facebook enable private groups and micro-influencers to grow in popularity and gain credibility.[204]

Fortunately, innovative tools and technologies to identify and mitigate the spread and impact of disinformation and manipulated content are likewise evolving.[205] Designed for various end users, these tools run the gamut from broad disinformation campaign analysis to individual self-help education.

[199] “Disinformation 101 Guide,” Disinfo Cloud, November 5, 2020, https://disinfocloud.com/blog/disinfo101-guide.

[200] Raina Davis, “Technology Factsheet: Deepfakes,” Harvard Kennedy School Belfer Center, Spring 2020, https://www.belfercenter.org/publication/technology-factsheet-deepfakes; Joan Donovan, “How memes got weaponized: A short history,” *MIT Technology Review*, October 24, 2019, https://www.technologyreview.com/2019/10/24/132228/political-war-memes-disinformation/; and Denise-Marie Ordway, “Information disorder: The essential glossary,” Harvard Kennedy School Shorenstein Center on Media, Politics and Public Policy, July 23, 2018, https://journalistsresource.org/studies/society/internet/information-disorder-glossary-fake-news.

[201] Joan Donovan and Claire Wardle, “Misinformation is Everybody’s Problem Now,” Social Science Research Council, August 6, 2020, https://items.ssrc.org/covid-19-and-the-social-sciences/mediated-crisis/misinformation-is-everybodys-problem-now/; and “Disinformation That Kills: The Expanding Battlefield Of Digital Warfare,” *CB Insights*, October 21, 2020, https://www.cbinsights.com/research/future-of-information-warfare/.

[202] “Tech Trends 2020,” Future Today Institute, accessed December 6, 2020, http://futuretodayinstitute.com/2020-tech-trends/.

[203] William Davies, “What’s wrong with WhatsApp,” *The Guardian*, July 2, 2020, https://www.theguardian.com/technology/2020/jul/02/whatsapp-groups-conspiracy-theories-disinformation-democracy; and Samuel Woolley, “Encrypted messaging apps are the future of propaganda,” Brookings Institution, May 1, 2020, https://www.brookings.edu/techstream/encrypted-messaging-apps-are-the-future-of-propaganda/.

[204] Eileen Brown, “Celebrity influencers on the wane: Most brands will choose micro-influencers in 2020,” *ZDNet*, March 20, 2020, https://www.zdnet.com/article/celebrity-influencers-on-the-wane-most-brands-will-choose-micro-influencers-in-2020/.

[205] “Tools Overview,” Disinfo Cloud, accessed December 12, 2020, https://disinfocloud.com/tools-overview/.

The following section provides an overview of some common categories of tools and technologies relevant to the counter-disinformation space. Many of these tools are featured on Disinfo Cloud, and some have gone through the Global Engagement Center's Technology Testbed, where they were assessed and validated against specific-use cases.[206]

Blockchain-based media authentication

Authenticating the provenance of media is an important step in countering disinformation, but it requires wide-scale adoption to be effective. Blockchain-based media authentication tools offer a potential solution. These tools maintain a decentralized, digital record of uploaded content, creating a "fingerprint," so content cannot be altered retroactively. This helps ensure the validity of original content and can provide a bulwark against claims of doctored media. Truepic, for example, can verify that an image or video hasn't been changed or edited and watermarks media with a time-stamp, geocode, and other metadata.[207] Similarly, Attestiv uses blockchain and AI technologies to track whether media content has been altered by adding "fingerprints" to a privately stored, distributed ledger with no identifiable data stored.[208] Once media has been added to the ledger, there is a clear chain of custody and record of authenticity. Artificial intelligence enhancements via perceptual fingerprinting can confirm validity even if the media has been re-saved in a different format or resolution.[209]

Dark web monitoring

While useful for activists evading repression and censorship, researchers estimate that 57 percent of live dark web sites host illicit activity, including activity that can facilitate disinformation campaigns.[210] Dark web monitoring tools can therefore be helpful in identifying and preparing for disinformation and propaganda activities that actors may later execute on the surface web. For example, Terbium Labs can identify ransomware for sale, stolen voter registration lists that may later be used in phishing attacks, or potential disinformation narratives being developed on the dark web.[211]

Recorded Future, another cybersecurity company specializing in the dark web, uncovered disinformation services on underground criminal forums run by criminal threat actors and nation-states.[212] According to Recorded Future, the disinformation services were easily accessible, costing anywhere from a few hundred to hundreds of thousands of dollars. Providers could create bulk social media accounts and publish content on media sources ranging from dubious websites to more reputable news outlets.

[206] Disinfo Cloud, accessed December 12, 2020., https://disinfocloud.com; and "Programs – Technology Engagement Team," U.S. Department of State, accessed December 12, 2020, https://www.state.gov/programs-technology-engagement-team/. The examples listed in this section represent just a sampling of the tools and technologies in the counter-disinformation space and do not indicate an endorsement by the Global Engagement Center or Park Advisors.

[207] Truepic website, accessed December 12, 2020, https://truepic.com/.

[208] Attestiv website, accessed December 12, 2020, https://attestiv.com/.

[209] Samuel Haig, "Attestiv CEO on Using DLT to Fight Fake News, Insurance Fraud, and Deep-Fakes," *Cointelegraph*, March 22, 2020, https://cointelegraph.com/news/attestiv-ceo-on-using-dlt-to-fight-fake-news-insurance-fraud-and-deep-fakes.

[210] Daniel Moore and Thomas Rid, "Cryptopolitik and the Darknet," *Global Politics and Strategy*, Volume 58, Issue 1, February 1, 2016, https://www.tandfonline.com/doi/full/10.1080/00396338.2016.1142085.

[211] Terbium Labs website, accessed December 12, 2020, https://terbiumlabs.com/; and "Election Security Primer: An Overview of the Most Pressing Threats Facing the 2020 U.S. Election," Terbium Labs, November 1, 2020, https://terbiumlabs.com/2020/11/01/election-security-primer-an-overview-of-the-most-pressing-threats-facing-the-2020-u-s-election/.

[212] Recorded Future website, accessed December 12, 2020, https://www.recordedfuture.com/; and "The Price of Influence: Disinformation in the Private Sector," Recorded Future, September 30, 2019, https://www.recordedfuture.com/disinformation-service-campaigns.

Using its DarkBlue Threat Intelligence Platform, Bluestone Analytics uncovered indications of disinformation, likely by Russian actors, originating on open web "news" sites deliberating posting to a dark web forum associated with the conspiracy group QAnon.[213]

Fact-checking

It has never been easier to create and publish content, but not all content is created equal. Fact-checking tools aggregate, analyze, and provide ratings on the credibility of information sources using various metrics to help users determine a website's reliability. NewsGuard, for example, employs a team of trained journalists and experienced editors to review and rate news and information websites based on nine journalistic criteria and then assigns a red or green rating, indicating the sites' credibility.[214]

Fact-checking tools are also useful for strengthening people's psychological resilience against misinformation and disinformation by raising awareness about the veracity of their information intake. For example, UK-based Full Fact has been checking claims made by politicians in the media and online for more than 10 years.[215] During that time, its team has built up an evidence base that can help people understand why misleading information arises, how it spreads, and who is responsible.[216] Another example is Trend Micro Check, a free tool created by cybersecurity company TrendMicro that helps users identify disinformation on popular social messaging apps such as Facebook Messenger, WhatsApp, and LINE. Users can add Trend Micro Check to their chat groups on various social messaging apps to check if a claim is false, or they can choose to send claims directly to Trend Micro Check. It is one of the top third-party verification chatbot services in Taiwan and recently expanded its services to Japan. [217]

Internet censorship circumvention

When states limit access to the internet as a means of control, internet censorship circumvention tools can help facilitate the continuous flow of information and promote free expression. One example, Psiphon, is a tool that operates in nearly 40 languages and helps 12 million active users connect to the internet every week.[218] Given its access and reach, Psiphon can also be used to aggregate and evaluate the presence of malware attacks globally, in almost real time. Psiphon's network has the capability to identify when a malware-infected device makes unrequested and unwanted contact with any malware command and control.[219] The presence of malware can make infected devices vulnerable to surveillance – often by a state-sponsored actor – and may be used to push disinformation and influence campaigns.[220] Psiphon can use its extensive data collection on regional internet connectivity patterns to provide aggregate data regarding malware attack patterns and can provide insight on internet blocking strategies that may be state-sponsored.

[213] "Russian Disinformation in the QAnon Boards," Bluestone Analytics, March 2, 2021, report on file with authors.

[214] NewsGuard website, accessed December 12, 2020, https://www.newsguardtech.com/ratings/rating-process-criteria.

[215] Full Fact website, accessed December 12, 2020, https://fullfact.org/.

[216] "Spotlight: Full Fact – Fighting the causes and consequences of bad information," Disinfo Cloud, May 6, 2020, https://disinfocloud.com/blog/spotlight-fullfact.

[217] "Identify Misinformation and Scams with Trend Micro Check," TrendMicro, November 17, 2020, https://www.trendmicro.com/en_us/research/20/k/trend-micro-check.html; and Audrey Tang and Sheau-Tyng Peng, "Dr. Message counters disinformation," PDIS, March 12, 2020, https://bit.ly/321Zica.

[218] "Spotlight: Psiphon," Disinfo Cloud, June 29, 2020, https://disinfocloud.com/blog/spotlight-psiphon.

[219] "Spotlight: Psiphon."

[220] "Spotlight: Psiphon."

Psychological resilience tools help audiences **increase their resilience to manipulated content by improving their media literacy and critical thinking skills.**

Other tools help identify new ways to bypass censorship. Geneva, developed by the Breakerspace Lab at the University of Maryland, is an experimental AI algorithm that analyzes country-level censorship technology and uses AI to learn and develop new censorship evasion strategies.[221] Researchers tested Geneva against the censorship technology of China, India, Iran, and Kazakhstan and found it able to suggest many known evasion strategies, as well as generate additional novel evasion approaches.

Manipulated information assessment

Some disinformation campaigns use fake profile pictures, bots, and other inauthentic content to provide the illusion of legitimacy. Manipulated information assessment tools use contextual clues to alert users to the potential that text, visuals, or audio may be manipulated. For example, Cyabra has developed AI-based technologies to identify malicious actors and determine the authenticity of interactions on social media platforms.[222] Their system examines around 250 metrics of social network users' behavior to assess whether an account is a regular user, a malicious bot, or a pseudo-human. WeVerify and OSoMe offer a number of open-source tools that are freely available to help users with disinformation analysis, such as analyzing rumors, bot detection, social network analysis, and analyzing memes and ads.[223]

Psychological resilience

While the onus should not be solely on the information consumer to identify and avoid disinformation, understanding how information can be misleading or manipulated is a valuable skill in today's information environment. Psychological resilience tools help audiences increase their resilience to manipulated content by improving their media literacy and critical thinking skills. They facilitate digital and media learning, quantifiably enhance cognitive performance, and can familiarize users with common disinformation tactics. For example, Harmony Square is a short, free-to-play online game in which players learn how disinformation is produced and spread.[224] Researchers found the game effective in reducing some of the harmful effects manipulative content can have on individuals.[225] Other games developed to educate players about disinformation include Troll Factory, which takes players through a week of "work" as a troll spreading disinformation for profit.[226]

Further examples of psychological resilience tools include flagging or listing false content to increase user familiarity with such content and help them better identify it. The online crowdsourced learning

[221] "Geneva: Evolving Censorship Evasion," censorship.ai, accessed December 12, 2020, http://geneva.cs.umd.edu/; and BreakerSpace website, University of Maryland, accessed December 12, 2020, https://breakerspace.cs.umd.edu/.

[222] Sarah Toth Stub, "Israeli cyber-sleuths hunt down fake news peddlers," *Times of Israel*, November 22, 2020, https://www.timesofisrael.com/spotlight/israeli-cyber-sleuths-hunt-down-fake-news-peddlers/.

[223] WeVerify website, accessed December 11, 2020, https://weverify.eu/; and Observatory on Social Media, Indiana University, accessed December 11, 2020, https://osome.iuni.iu.edu/; and "Spotlight: Using WeVerify and OSoMe Tools to Track COVID-19 Misinformation," Disinfo Cloud, April 22, 2020, https://disinfocloud.com/blog/spotlight-weverify-osome.

[224] "Spotlight: Harmony Square – Exposing disinformation tactics and techniques," Disinfo Cloud, November 10, 2020, https://disinfocloud.com/blog/spotlight-harmonysquare.

[225] Jon Roozenbeek and Sander van der Linden, "Breaking Harmony Square: A game that 'inoculates' against political misinformation," *MisinfoReview*, Harvard Kennedy School, November 6, 2020, https://misinforeview.hks.harvard.edu/article/breaking-harmony-square-a-game-that-inoculates-against-political-misinformation/.

[226] Troll Factory website, accessed December 11, 2020, https://trollfactory.yle.fi/.

platform, Mind Over Media, an initiative from the University of Rhode Island's Media Education Lab, hosts more than 3,500 examples of current propaganda from over 40 countries, along with a suite of nine lesson plans suitable for use in high school, college, and with adult learners.[227]

Social listening

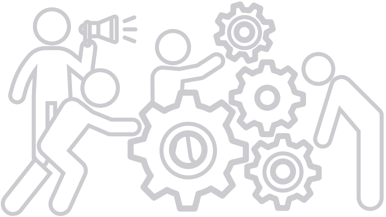
Tracking key narratives, popular influencers, and what resonates or not is crucial in the counter-disinformation space. Social listening tools help users understand the online information environment by monitoring information from social media channels. Platforms like Zignal Labs and Meltwater can assist in understanding how information is shared or spread via social media, help identify bots and trolls, and offer insight into the effectiveness of a media campaign.[228] Other tools like Yonder can also be used for early warning to alert users to suspicious online activity to identify and address disinformation campaigns.[229] Many of these platforms also allow users to visualize trends and emerging topics to understand their information environments. For example, a data collection campaign by IST Research captured more than 400,000 documents related to Russian-language online conversations around COVID-19 posted on four different platforms.[230] Their analysis of the data revealed that the Russian news community and social media trolls exploited the pandemic to disseminate inflammatory narratives and outright falsehoods.

[227] Mind Over Media website, accessed December 11, 2020, https://propaganda.mediaeducationlab.com/.

[228] Zignal Labs website, accessed December 11, 2020, https://zignallabs.com/; and "Media Monitoring," Meltwater, accessed December 11, 2020, https://www.meltwater.com/en/media-monitoring.

[229] Yonder website, accessed December 11, 2020, https://www.yonder-ai.com/; and Taylor Hatmaker, "Coronavirus conspiracies like that bogus 5G claim are racing across the internet," *TechCrunch*, April 10, 2020, https://techcrunch.com/2020/04/10/coronavirus-5g-covid-19-conspiracy-theory-misinformation/?guccounter=1.

[230] IST Research website, accessed December 14, 2020, https://www.istresearch.com/; and "Spotlight: IST Research – Russian COVID-19 Narratives," Disinfo Cloud, April 3, 2020, https://disinfocloud.com/blog/spotlight-istresearch.



5. Conclusion

As evolving and emerging technologies provide new ways to identify, mitigate, and counter disinformation, they also offer new ways for disinformation campaigns, tactics, and actors to evolve and adapt. The cat-and-mouse game continues apace, further cementing disinformation as an enduring problem. What becomes clear is that disinformation is not a problem to be solved but rather one that can only be managed. So, who should do the managing?

While everyone has a role in managing disinformation, the responsibility cannot and should not be distributed evenly. Primary responsibility lies with the social media and tech platforms that need to act more consistently and transparently in their algorithm and feature developments, as well as their content moderation, regardless of current events or public pressure. Other roles flow from there. Academia, civil society, media, and others can better address disinformation if provided more data from, or are afforded the opportunity for meaningful collaboration with, social media and tech platforms When governments get involved through legislation or regulation, they must carefully consider the free speech and freedom of expression implications.

Then there's the information consumer. Thanks to the numerous online tools that help people discern truthful information, it's never been easier to verify the veracity of content online. Yet this assumes that the average person is motivated to seek out the tools they need to ensure they avoid consuming and sharing disinformation. Research on human behavior and cognitive bias is clear that this is not the case.[231] The onus should therefore not be on the information consumer to actively resist disinformation but should instead shift to ensure people can passively resist disinformation. This means the pendulum of responsibility shifts back to the social media and tech platforms to integrate content verification into common platforms and tools as a default feature, so users are alerted to questionable content at the exact moment they encounter it.

To reiterate, more proactive action from the social media and tech platforms, while helpful, will not solve the challenges associated with disinformation. However, their leadership, combined with significant collaborations across industries, provides the best opportunity available to manage this ever-evolving problem.

[231] Christina Nemr and William Gangware, "Weapons of Mass Distraction: Foreign State-Sponsored Disinformation in the Digital Age," *Park Advisors*, March 2019, https://www.park-advisors.com/disinforeport.

DISINFOCLOUD.COM