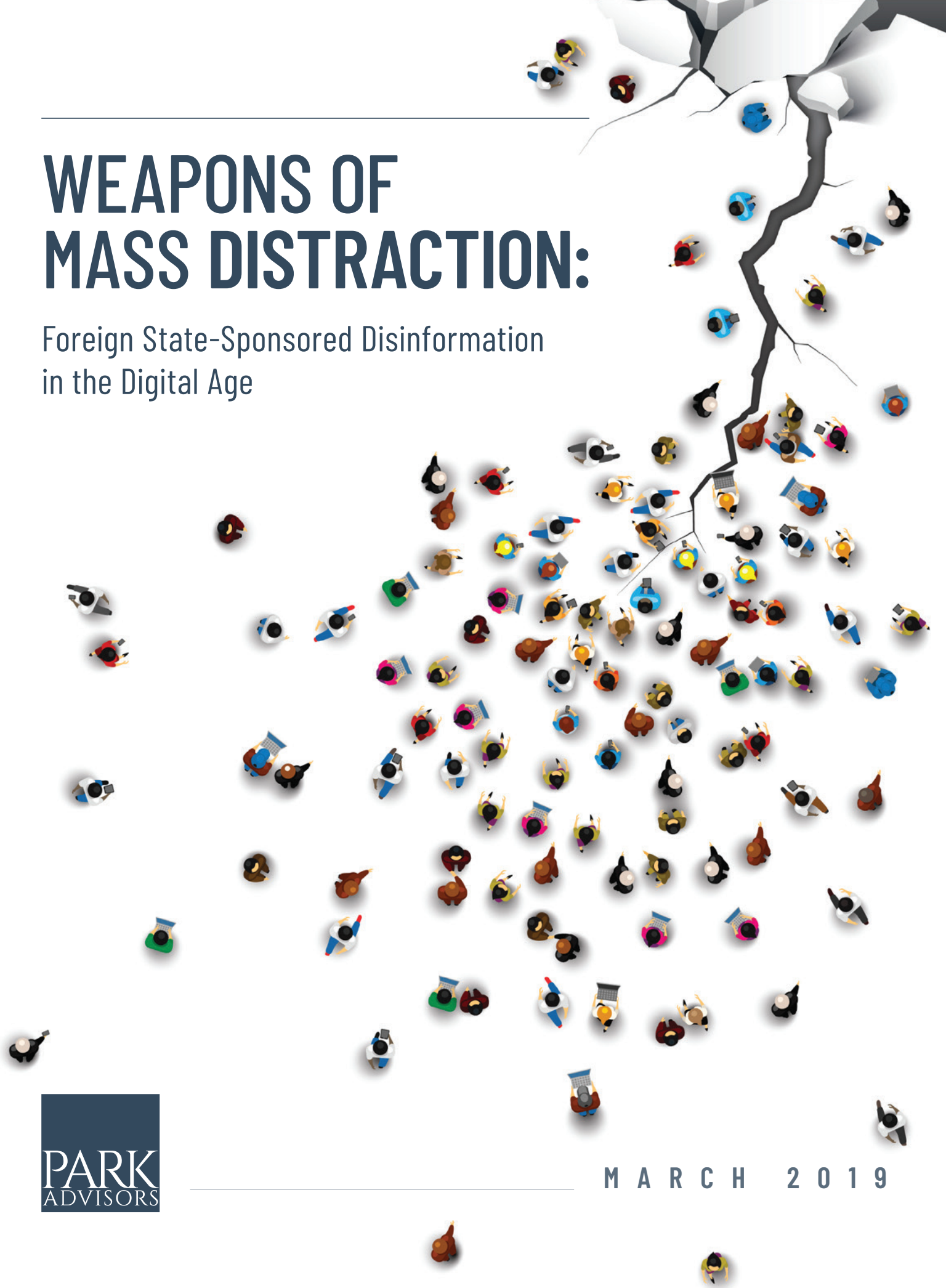
WEAPONS OF MASS DISTRACTION:

Foreign State-Sponsored Disinformation in the Digital Age



PARK ADVISORS

MARCH 2019

Authored by

Christina Nemr and William Gangware

Acknowledgements

The authors are grateful to the following subject matter experts who provided input on early drafts of select excerpts: Dr. Drew Conway, Dr. Arie Kruglanski, Sean Murphy, Dr. Alina Polyakova, and Katerina Sedova. The authors also appreciate the contributions to this paper by Andrew Rothgaber and Brendan O'Donoghue of Park Advisors, as well as the editorial assistance provided by Rhonda Shore and Ryan Jacobs.

This report was produced with support from the US Department of State's Global Engagement Center. Any views expressed in this report are those of the authors and do not necessarily reflect the views of the US State Department, Park Advisors, or its subject matter expert consultants. Any errors contained in this report are the authors' alone.

0. Table of Contents

1. Introduction and contextual analysis

On July 12, 2014, viewers of Russia's main state-run television station, Channel One, were shown a horrific story. Five months prior, the Russian military had pushed its way into neighboring Ukraine, and Channel One had been covering the political and military action on the ground. The July 12 story, however, was particularly dramatic.

That day, Channel One reporters interviewed a woman at a refugee camp near the Russian border, who claimed to witness a squad of Ukrainian soldiers nail a three-year-old boy to a post in her town square. The soldiers had tortured the boy to death over a period of hours, before tying his mother to the back of a tank and dragging her through the square. [1]

Channel One never questioned the woman's story. But at least one independent Russian journalist found the tale so unbelievable that he visited the alleged site to investigate. Finding no evidence that this atrocity had ever occurred, he interviewed one resident in the town square, the supposed site of the crime. "This," the resident said, "is the first I'm hearing of it."

So where did the story come from? Within a few days, the journalist and others traced the story back to a political scientist with ties to the Kremlin. Days before the shocking Channel One report, this man had posted a similar tale on Facebook, with nearly identical details. By the time the Kremlin connection was uncovered, however, the damage was done: not only had the Channel One report run on television, but the viral story was now reaching a much broader audience on social media.

The false crucifixion story was but one example of Kremlin-backed disinformation deployed during Russia's annexation of Crimea. In subsequent years, similar tactics would again be unleashed by the Kremlin on other foreign adversaries, including the United States during the lead-up to the 2016 presidential election.

1. See "State-Run News Station Accused of Making Up Child Crucifixion," *The Moscow Times*, 14 July 2014, https://themoscowtimes.com/news/state-run-news-station-accused-of-making-up-child-crucifixion-37289; and Arkady Ostrovsky, "Putin's Ukraine Unreality Show," *Wall Street Journal*, 28 July 2014, https://www.wsj.com/articles/arkady-ostrovsky-putins-ukraine-unreality-show-1406590397; and Andrew Higgins, "Fake News, Fake Ukrainians, How a Group of Russians Tilted a Dutch Vote," *New York Times*, 16 Feb 2017, https://www.nytimes.com/2017/02/16/world/europe/russia-ukraine-fake-news-dutch-vote.html.

Yet the use of modern-day disinformation does not start and end with Russia. A growing number of states, in the pursuit of geopolitical ends, are leveraging digital tools and social media networks to spread narratives, distortions, and falsehoods to shape public perceptions and undermine trust in the truth.

If there is one word that has come to define the technology giants and their impact on the world, it is "disruption." The major technology and social media companies have disrupted industries ranging from media to advertising to retail. However, it is not just the traditional sectors that these technologies have upended. They have also disrupted another, more insidious trade – disinformation and propaganda.

A growing number of states, **in the pursuit of geopolitical ends**, are leveraging digital tools and social media networks to spread narratives, distortions, and **falsehoods to shape public perceptions and undermine trust in the truth**.

The proliferation of social media platforms has democratized the dissemination and consumption of information, thereby eroding traditional media hierarchies and undercutting claims of authority. The environment, therefore, is ripe for exploitation by bad actors. Today, states and individuals can easily spread disinformation at lightning speed and with potentially serious impact.

There are significant vulnerabilities in the information ecosystem that foreign state-sponsored actors can exploit, and they revolve around three primary, interconnected elements:

1. *The **medium*** – the platforms on which disinformation flourishes;
2. *The **message*** – what is being conveyed through disinformation; and
3. *The **audience*** – the consumers of such content.

The first two elements, the medium and the message, operate hand in hand. Social media and news platforms are designed to deliver information to mass audiences quickly, optimizing for viral content that generates clicks and thus revenue. As a consequence, they are inherently vulnerable to sensationalist disinformation that seeks to catch the eye and be shared.[2]

The messages conveyed through disinformation range from biased half-truths to conspiracy theories to outright lies. The intent is to manipulate popular opinion to sway policy or inhibit action by creating division and blurring the truth among the target population.

Unfortunately, the most useful emotions to create such conditions – uncertainty, fear, and anger – are the very characteristics that increase the likelihood a message will go viral. Even when disinformation first appears on fringe sites outside of the mainstream media, mass coordinated action that takes advantage of platform business models reliant upon clicks and views helps ensure greater audience penetration.[3] Bot networks consisting of fake profiles amplify the message and create the illusion of high activity and popularity across multiple platforms at once, gaming recommendation and rating algorithms.

2. Information Society Project at Yale Law School and the Floyd Abrams Institute for Freedom of Expression, "Fighting Fake News (Workshop Report)," 2017, https://law.yale.edu/system/files/area/center/isp/documents/fighting_fake_news_-_workshop_report.pdf.
3. "Connecting the bots: Researchers uncover invisible influence on social media," University of Georgia, 30 May 2017, https://www.sciencedaily.com/releases/2017/05/170530095910.htm.

On average, **a false story reaches 1,500 people six times more quickly than a factual story**. This is true of false stories about any topic, but **stories about politics are the most likely to go viral**.

Research shows that these techniques for spreading fake news are effective. On average, a false story reaches 1,500 people six times more quickly than a factual story.[4] This is true of false stories about any topic, but stories about politics are the most likely to go viral.[5]

For all that has changed about disinformation and the ability to disseminate it, arguably the most important element has remained the same: the audience. No number of social media bots would be effective in spreading disinformation if the messages did not exploit fundamental human biases and behavior. People are not rational consumers of information. They seek swift, reassuring answers and messages that give them a sense of identity and belonging.[6] The truth can be compromised when people believe and share information that adheres to their worldview.

The problem of disinformation is therefore not one that can be solved through any single solution, whether psychological or technological. An effective response to this challenge requires understanding the converging factors of technology, media, and human behaviors.

The following interdisciplinary review attempts to shed light on these converging factors, and the challenges and opportunities moving forward.

4. Robinson Meyer, "The Grim Conclusions of the Largest-Ever Study of Fake News," *The Atlantic*, 08 March 2018, https://www.theatlantic.com/technology/archive/2018/03/largest-study-ever-fake-news-mit-twitter/555104/.
5. Meyer, "The Grim Conclusions," *The Atlantic*.
6. Daniele Anastasion, "The Price of Certainty," *New York Times*, 01 November 2016, https://www.nytimes.com/2016/11/01/opinion/the-price-of-certainty.html.

2. How do we define disinformation?

Several terms and frameworks have emerged to describe information that misleads, deceives, and polarizes. The most popular of these terms are misinformation and disinformation, and while they are sometimes used interchangeably, researchers agree they are separate and distinct.

Misinformation is generally understood as the *inadvertent* sharing of false information that is not intended to cause harm.[7] Disinformation, on the other hand, is widely defined as the *purposeful* dissemination of false information intended to mislead or harm.[8] Although a straightforward definition, it can be difficult to ascribe precise parameters to disinformation. For example, disinformation is not always composed of fabrications. It can consist of true facts, pieced together to portray a distorted view of reality.[9]

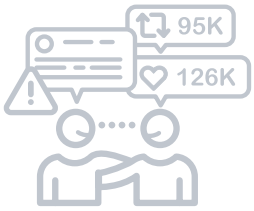
To understand the disinformation environment, it is useful to dissect the different elements it encompasses.[10] Disinformation can include authentic material used in a deliberately wrong context to make a false connection, such as an authentic picture displayed with a fake caption. It can take the form of fake news sites or ones that are deliberately designed to look like well-known sites. Disinformation can further include outright false information, shared through graphics, images, and videos. It can also take the form of manipulated image and video content, where controversial elements are photoshopped into innocuous contexts to evoke anger or outrage.

7. Hossein Derakhshan and Clair Wardle, "Information Disorder: Definitions" in *Understanding and Addressing the Disinformation Ecosystem*, Annenberg School for Communications workshop, 15-16 December 2017, pp. 5-12, https://firstdraftnews.org/wp-content/uploads/2018/03/The-Disinformation-Ecosystem-20180207-v2.pdf.
8. Issue Brief: "Distinguishing Disinformation from Propaganda, Misinformation, and "Fake News"," National Endowment for Democracy, 17 October 2017, https://www.ned.org/issue-brief-distinguishing-disinformation-from-propaganda-misinformation-and-fake-news/.
9. See note 8$_2$; and Don Fallis, "The Varieties of Disinformation" in *The Philosophy of Information Quality* [Boston: Northeastern University, 2014], pp.135-161, https://www.researchgate.net/publication/321554157_The_Philosophy_of_Information_Quality; and Alice E. Marwick,_ "Why Do People Share Fake News? A Sociotechnical Model of Media Effects," *Georgetown Law Technology Review, 2018,* https://www.georgetownlawtechreview.org/wp-content/uploads/2018/07/2.2-Marwick-pp-474-512.pdf.
10. For more, see Claire Wardle, "6 Types of Misinformation Circulated This Election Season," *Columbia Journalism Review*, 18 November 2016, https://www.cjr.org/tow_center/6_types_election_fake_news.php; and Fallis, "The Varieties of Disinformation," https://www.researchgate.net/publication/278692847_The_Varieties_of_Disinformation.

A note on terms

This interdisciplinary review is meant to explore the topic of disinformation, understanding it as a term distinct from misinformation. However, the literature on the topic overwhelmingly uses the terms *misinformation*, *disinformation*, and even *fake news* and *propaganda* interchangeably. This review therefore uses the terms as stated in the specific literature to adhere to the spirit of the research.

Furthermore, while this review is focused on the ways state actors use disinformation to further geopolitical goals, the analysis of disinformation contained herein encompasses additional contexts outside of geopolitics in the interest of presenting a thorough review.



3. What psychological factors drive vulnerabilities to disinformation and propaganda?

Disinformation succeeds, in part, because of psychological vulnerabilities in the way people consume and process information. Indeed, experts on a 2018 *Newsgeist* panel – a gathering of practitioners and thinkers from journalism, technology, and public policy – identified a number of psychological features that make disinformation so effective with audiences. These features include how disinformation plays to emotions and biases, simplifies difficult topics, allows the audience to feel as though they are exposing truths, and offers identity validation.[11]

The following section reviews some of these psychological factors and their implications for the appeal and persistence of disinformation.

The need to belong

A large body of research shows that people desire social belonging, such as inclusion within a community, and the resulting identity that accompanies such belonging.[12] Indeed, the research indicates that this need for belonging is a fundamental human motivation that dictates most interpersonal behavior.[13] These motivations play out in real time online, often with drastic effect. For better or worse, the internet and social media have facilitated the ability to seek out and find a community that contributes to a person's sense of belonging. In particular, research shows that social media can provide positive psychosocial well-being, increase social capital, and even enable offline social interactions.[14]

11. "What's disinformation doing "right" – and what can newsrooms learn from it?," NiemanLab, 02 November 2018, http://www.niemanlab.org/2018/11/whats-disinformation-doing-right-and-what-can-newsrooms-learn-from-it/.
12. Gregory M. Walton and Geoffrey L Cohen, "A Question of Belonging: Race, Social Fit, and Achievement," *Journal of Personality and Social Psychology*, Vol. 92, No. 1, (2007), pp. 82-96, http://lmcreadinglist.pbworks.com/f/Walton+%26+Cohen+(2007).pdf.
13. Roy F. Bauermeister and Mark R. Leary, "The Need to Belong: Desire for Interpersonal Attachments as a Fundamental Human Motivation," *Psychological Bulletin*, 1995, http://persweb.wabash.edu/facstaff/hortonr/articles%20for%20class/baumeister%20and%20leary.pdf.
14. Elizabeth A. Vincent, "Social Media as an Avenue to Achieving Sense of Belonging Among College Students," Vistas Online, American Counseling Association, 2016, https://www.counseling.org/docs/default-source/vistas/social-media-as-an-avenue.pdf?sfvrsn=f684492c_8.

So how does this apply to the resonance of disinformation and propaganda? In their desire for social belonging, people are interested in consuming and sharing content that connects with their own experiences and gives shape to the identity and status they want to project.[15] Status-seeking and identity projection therefore play a consequential role in motivating people to share stories.[16]

Additionally, the nature of social media amplifies the choices people make in service to social identity and belonging because its activity takes place in a public forum.[17]

The implications of social identity and belonging further extend to the sources people trust when consuming information. Research shows that individuals depend on their social networks as trusted news sources and are more likely to share a post if it originates from a trusted friend.[18] This can increase susceptibility to disinformation if one's network is prone to sharing unverified or low-quality information.

The sheer volume of content is another important factor enabling disinformation. **Six thousand tweets are sent every second** and **Facebook hosts 300 million new photos every day**.

Dealing with the firehose of information

The sheer volume of content is another important factor enabling disinformation. Six thousand tweets are sent every second and Facebook hosts 300 million new photos every day.[19] Research reveals that individuals are ill-equipped to assess and cope with such volume, leading them to quickly discard what they perceive as irrelevant or unwanted information.[20] The volume of information, combined with people's limited attention spans, also makes it difficult to discern between high- and low-quality information, creating an environment in which disinformation has the potential to go viral.[21]

How do people then respond to all this information? Although people like to believe they are rational consumers of information, decades of psychological research has demonstrated the limitations of the human brain when it comes to information intake, particularly in contexts of uncertainty and anxiety.[22] Humans are generally averse to uncertain and ambiguous situations, leading them to seek quick, definitive answers.[23]

15. See Douglas Rushkoff, David Pescovitz, and Jake Dunagan, "The Biology of Disinformation," Institute for the Future, 2018, http://www.iftf.org/fileadmin/user_upload/images/ourwork/digintel/IFTF_biology_of_disinformation_062718.pdf; and Alice E. Marwick, "Why Do People Share Fake News? A Sociotechnical Model of Media Effects."
16. Rushkoff et al, "The Biology of Disinformation," Institute for the Future, 2018.
17. Natalie Jomini Stroud, Emily Thorson, and Dannagal Young, "Making Sense of Information and Judging its Credibility" in *Understanding and Addressing the Disinformation Ecosystem*, Annenberg School for Communications workshop,15-16 December 2017, pp. 45 -50 https://firstdraftnews.org/wp-content/uploads/2018/03/The-Disinformation-Ecosystem-20180207-v2.pdf.
18. Marwick, "Why Do People Share Fake News? A Sociotechnical Model of Media Effects."
19. Jomini Stroud, Thorson, Young, "Making Sense of Information and Judging its Credibility."
20. Xiaoyan Qiu et al, "Limited individual attention and online virality of low-quality information," *Nature Human Behavior*, 26 June 2017, https://www.nature.com/articles/s41562-017-0132.
21. Qiu et al, "Limited individual attention and online virality of low-quality information," ***Nature Human Behavior***.
22. Donna M. Webster and Arie W. Kruglanski, "Cognitive and Social Consequences of the Need for Cognitive Closure," *European Review of Social Psychology*, 15 April 2011, https://www.tandfonline.com/doi/abs/10.1080/14792779643000100?journalCode=pers20.
23. Maria Konnikova, "Why We Need Answers," *The New Yorker*, 30 April 2013, https://www.newyorker.com/tech/annals-of-technology/why-we-need-answers.

Arie Kruglanski, a psychology professor at the University of Maryland, defines this phenomenon as the need for *cognitive closure*, or certainty in an uncertain world.[24] Though seeking answers in the face of uncertainty is a natural human motivation, further research indicates that the answers upon which people seize can be so clear-cut as to be detrimental, specifically when such answers promote black and white thinking and disallow other viewpoints.[25] In particular contexts, this can create the conditions conducive to the extremism and polarization that allows disinformation to flourish.[26]

Similar to cognitive closure, the literature has identified other cognitive biases that dictate how people take in and interpret information to help them make sense of the world. For example, *selective exposure* leads people to prefer information that confirms their preexisting beliefs, while *confirmation bias* makes information consistent with one's preexisting beliefs more persuasive.[27] These biases interact with, and complement, two other types of bias: *motivated reasoning* and *naïve realism*.[28]

While confirmation bias leads individuals to seek information that fits their current beliefs, motivated reasoning is the tendency to apply higher scrutiny to unwelcome ideas that are inconsistent with one's ideas or beliefs.[29] In this way, people use motivated reasoning to further their quest for social identity and belonging.

Further entrenching the effects of these biases, the research shows that naïve realism plays an important role during the intake and assessment of information. Naïve realism leads individuals to believe that their perception of reality is the only accurate view, and that those who disagree are simply uninformed or irrational.[30]

These cognitive biases show that although individuals may believe their convictions are based on higher principles, in reality people process moral-political statements as preferences as opposed to hard facts.[31] Given that preferences change throughout one's life, particular convictions may also change in the right context of persuasion, attitude change, or social network. This is especially true of convictions that are more controversial or subject to mixed public consensus, like politics.[32]

Cognitive limitations in an online jungle

So how do these cognitive biases play out in the social media sphere? A 2016 study of news consumption on Facebook examined 376 million users and 920 news outlets to answer this question. They found that

24. Arie W. Kruglanski and Donna M. Webster, "Motivated Closing of the Mind: "Seizing" and "Freezing"," National Center for Biotechnology Information, April 1996, https://www.ncbi.nlm.nih.gov/pubmed/8637961.
25. D. Webber et al, "The Road to Extremism: Field and Experimental Evidence that Significance Loss-Induced Need for Closure Fosters Radicalization," *US National Library of Medicine*, 04 September 2017, https://www.ncbi.nlm.nih.gov/pubmed/28872332.
26. Webber et al, "The Road to Extremism: Field and Experimental Evidence that Significance Loss-Induced Need for Closure Fosters Radicalization."
27. David M. J. Lazer et al, "The Science of Fake News," *Science*, 09 March 2018, http://science.sciencemag.org/content/359/6380/1094/tab-pdf.
28. Drew Calvert, "The Psychology Behind Fake News," *Kellogg Insight*, 06 March 2017, https://insight.kellogg.northwestern.edu/article/the-psychology-behind-fake-news.
29. Gary Marcus, "How Does the Mind Work? Insights from Biology," *Topics in Cognitive Science*, 17 November 2008, http://www.psych.nyu.edu/gary/marcusArticles/Marcus%202009%20topics.pdf.
30. Calvert, "The Psychology Behind Fake News."
31. Calvert, "The Psychology Behind Fake News."
32. Calvert, "The Psychology Behind Fake News."

Users are more active in sharing unverified rumors than they are in later sharing that these rumors were either debunked or verified. **The veracity of information therefore appears to matter little.**

users tend to confine their attention to a limited set of pages, seeking out information that aligns with their views and creating polarized clusters of information sharing.[33]

In another study, researchers assessed 330 rumor threads on Twitter associated with nine newsworthy events, such as the December 2014 Sydney hostage siege and the January 2015 Charlie Hebdo shooting in Paris, to understand how people interact with rumors on social media. Their analysis determined that users are more active in sharing unverified rumors than they are in later sharing that these rumors were either debunked or verified.[34] The veracity of information therefore appears to matter little. A related study found that even after individuals were informed that a story had been misrepresented, more than a third still shared the story.[35]

In addition to highlighting the limitations of human cognition, the research also points to declining trust in the public sphere. Richard Fletcher and Rasmus Nielsen from the University of Oxford argue that disinformation must be analyzed in the context of other factors, including declining trust in news media and increasing skepticism of online information, which has been exacerbated by clickbait and advertisements that masquerade as news.[36]

In a complementary study, researchers found that participants who perceived the media and the word "news" negatively were less likely than others to identify a fake headline and less able to distinguish news from opinion or advertising.[37]

The varying levels of trust in the media have implications for efforts to validate the veracity of news. For example, tagging social media posts as "verified" may work well in environments where trust in news media is relatively high (such as Spain or Germany), but this approach may be counterproductive in countries where trust in news media is much lower (like Greece).[38]

Doubling down online

Given the human motivations that drive online behavior, researchers contend that it is more likely that polarization exacerbates fake news, rather than fake news exacerbating polarization.[39]

33. Ana Lucia Schmidta et al, "Anatomy of news consumption on Facebook," *Proceedings of the National Academy of Sciences*, 21 March 2017, http://www.pnas.org/content/pnas/114/12/3035.full.pdf.
34. Arkaitz Zubiaga et al, "Analysing How People Orient to and Spread Rumors in Social Media by Looking at Conversational Threads," PLOS ONE 11 (3): e0150989, 04 March 2016, https://doi.org/10.1371/journal.pone.0150989.
35. Laura Hazard Owen, "Americans may appreciate knowing when a news story is suspect, but more than a third will share that story anyway," *Nieman Lab*, 29 June 2018, http://www.niemanlab.org/2018/06/americans-may-appreciate-knowing-when-a-news-story-is-suspect-but-more-than-a-third-will-share-that-story-anyway/.
36. Richard Fletcher and Rasmus Nielsen, "People Don't Trust News Media – and this is Key to the Global Misinformation Debate" in *Understanding and Addressing the Disinformation Ecosystem*, Annenberg School for Communications workshop,15-16 December 2017, pp. 13-17, https://firstdraftnews.org/wp-content/uploads/2018/03/The-Disinformation-Ecosystem-20180207-v2.pdf.
37. "How the Public, News Sources, and Journalists Think about News in Three Communities," News Co/Lab at Arizona State University in collaboration with the Center for Media Engagement at The University of Texas at Austin, 2018, https://mediaengagement.org/wp-content/uploads/2018/11/How-the-Public-News-Sources-and-Journalists.pdf.
38. Fletcher and Nielsen, "People Don't Trust News Media – and this is Key to the Global Misinformation Debate."
39. Calvert, "The Psychology of Fake News."

Humans react the same way to **undesirable information** as they do when facing a dangerous animal – **fight or flight**.

People's propensity toward "us versus them" tribalism applies just as much to the information they consume.

What, then, can be done to reduce polarization online? The literature highlights a number of challenges.

In an effort to avoid echo chambers, some have advocated for increasing online communities' exposure to different viewpoints. However, one study that attempted this approach found it to be not just ineffective, but counterproductive.[40] The study identified a large sample of Democrats and Republicans on Twitter and offered financial incentives to follow a bot that exposed the participants to messages of opposing political ideologies. The results were surprising: Republicans who followed the liberal bot became substantially more conservative, while Democrats who followed the conservative bot became slightly more liberal.

The study offers a cautionary tale for future efforts to reduce polarization online. The observed backfire effect may be explained by complementary research, which found that acknowledging unwelcome facts about controversial issues can be threatening.[41] Humans react the same way to undesirable information as they do when facing a dangerous animal – fight or flight.[42] To deal with the threat, people double down to defend their previously held beliefs or shun the new information rather than amend their views.[43]

Given this ingrained resistance to new ideas, can people change their minds? The jury is still out. The ability of individuals to adjust their perceptions after being shown corrected information may vary based on their cognitive ability.[44] One study, in which individuals were shown corrections to misinformation, found that individuals with low cognitive ability less frequently adjusted their viewpoints than those with high cognitive ability.[45] A similar study showed that an audience's level of cognitive activity is likely to predict the persistence of misinformation and effectiveness of a correction.[46]

The resonance of disinformation and why it is difficult to debunk

While some have argued for an increase in fact-checking or debunking efforts to counter disinformation, the literature is again mixed on the effectiveness of such approaches.

40. Christopher A. Bail et al, "Exposure to Opposing Views on Social Media Can Increase Political Polarization," *Proceedings of the National Academy of Sciences*, September 2018, http://www.pnas.org/content/115/37/9216.
41. Brendan Nyhan and Jason Reifler, "Misinformation and Fact-Checking: Research Findings from Social Science," New America Foundation, February 2012, https://www.dartmouth.edu/~nyhan/Misinformation_and_Fact-checking.pdf.
42. Arthur Lupia, "Communicating Science in Politicized Environments," *Proceedings of the National Academy of Sciences*, 20 August 2013, http://media.wix.com/ugd/fa8393_6973c3639e3c4bdfa2908cab10587cf4.pdf.
43. Nyhan and Reifler, "Misinformation and Fact-Checking: Research Findings from Social Science."
44. Jonas De Keersmaecker and Arne Roets, "'Fake news': Incorrect, but hard to correct. The role of cognitive ability on the impact of false information on social impressions," *Intelligence, Volume 65*, November 2017, pp. 107-110, https://doi.org/10.1016/j.intell.2017.10.005.
45. De Keersmaecker and Roets, "'Fake news': Incorrect, but hard to correct."
46. Man-pui Sally Chan et al, "Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation," *Psychological Science*, 12 September 2017, https://journals.sagepub.com/doi/full/10.1177/0956797617714579.

Disinformation is difficult to correct because of how it resonates cognitively and emotionally with its intended audience and how humans form beliefs based on the information they have at hand. This phenomenon is referred to as *belief perseverance*, which is the inability of people to change their minds even after being shown new information.[47] Facts can matter little in the face of strong social and emotional dynamics that relate to one's personal identity and sense of community.[48]

Others have advocated for increasing media literacy and making social media users more critical consumers of content. However, researchers have found that efforts to boost critical thinking may be of limited use or may have the counterproductive effect of leading individuals to doubt the credibility of news in general.[49] Research even indicates that many consumers of disinformation already perceive themselves as critical thinkers who are challenging the status quo.[50] The literature makes explicit that consumers are not well-equipped to identify certain messages as false or misleading, nor should the onus be on them to do so.

To underscore these challenges, one study analyzed the effectiveness of messages meant to reduce misinformation about the links between vaccines and autism. It found that while the messages occasionally reduced the belief that vaccines cause autism, they did not affect the behavior or the intent to vaccinate among parents who had the most negative attitudes on vaccines.[51] A similar study found the same phenomenon among skeptics of climate-change research. Evidence that countered their emotion-based beliefs did not fundamentally change those beliefs.[52]

Considering these challenges, are there any approaches to fact-checking that might work?

According to the researchers behind the vaccine study, one way to correct misinformation is by providing an alternate causal explanation to displace inferences made from the false information. However, other research casts doubt on how well even a thorough debunking statement will fare. A study found that debunking was ineffective when the consumer could generate competing explanations supporting the misinformation. Furthermore, a debunking message that simply identified misinformation as incorrect without offering corrective information was similarly ineffective. Even when a detailed debunking message included corrective information, the debunking did not always reduce participants' belief in the misinformation.[53]

47. Brendan Nyhan and Jason Reifler, "Displacing Misinformation about Events: An Experimental Test of Causal Corrections," *Journal of Experimental Political Science*, 01 April 2015, https://doi.org/10.1017/XPS.2014.22.
48. Calvert, The Psychology Behind Fake News," https://insight.kellogg.northwestern.edu/article/the-psychology-behind-fake-news; and Brendan Nyhan et al, "Effective Messages in Vaccine Promotion: A Randomized Trial," *Pediatrics*, April 2014, http://pediatrics.aappublications.org/content/133/4/e835.
49. Calvert, The Psychology Behind Fake News."
50. Deen Freelon, "Personalized Information Environments and Their Potential Consequences for Disinformation" in *Understanding and Addressing the Disinformation Ecosystem*, Annenberg School for Communications workshop,15-16 December 2017, pp. 38-44, https://firstdraftnews.org/wp-content/uploads/2018/03/The-Disinformation-Ecosystem-20180207-v2.pdf.
51. Brendan Nyhan et al, "Effective Messages in Vaccine Promotion: A Randomized Trial," *Pediatrics*, April 2014, http://pediatrics.aappublications.org/content/133/4/e835.
52. Paul Thagard and Scott Findlay, "Changing Minds About Climate Change: *Belief Revision, Coherence, and Emotion*," in Belief Revision Meets Philosophy of Science, eds. Eric J. Olsson and Sebastian Enqvis [Netherlands: Springer, 03 November 2010], http://cogsci.uwaterloo.ca/Articles/thagard.climate.2011.pdf.
53. Chan et al, "Debunking: A Meta-Analysis of the Psychological Efficacy of Messages Countering Misinformation."

Compounding matters is the fact that repeating a false claim can increase its believability.[54] In studies where participants rated statements on their truthfulness, results showed that repetition increased participants' perceptions of the truthfulness of false statements even when participants knew these statements were false.[55] Given that individuals are susceptible to familiarity of information, the repetition of verifiably false statements can decrease the power of individual knowledge and reasoning.[56] This finding has further implications for fact-checking: a fact-checker must repeat a false statement to debunk it, thereby inadvertently increasing the repetition and potential believability of the false claim.[57]

Another study showed that when consumers of fake news were presented with a fact-check, **they almost never read it**.

Other literature reveals that the nature of how one relates to misperceptions impacts the likelihood of accepting or rejecting corrections. For example, misperceptions tied to salient or controversial issues, particularly those that matter highly to an individual's concept of self, are highly persistent and hard to correct. Conversely, it may be easier to counter misperceptions on topics where people do not have strongly held beliefs.[58]

There is, however, a social element to fact-checking that can encourage more accurate information. For example, if there are strong social connections between individuals who generate false content and individuals who fact-check this content, the former are more likely to correct their false statements.[59] Unfortunately, because corrected statements are often less read than the misleading original statement, it is unclear how effective such corrections are.[60] Another study showed that when consumers of fake news were presented with a fact-check, they almost never read it.[61]

Fighting fire with fire

So, what strategies might work to counter disinformation? Recent research is more positive regarding potential approaches.

54. Adam J. Berinsky, "Rumors and Healthcare Reform: Experiments in Political Misinformation," *British Journal of Political Science*, April 2017, https://doi.org/10.1017/S0007123415000186.
55. Lisa K. Fazio et al, "Knowledge does not protect against illusory truth," *Journal of Experimental Psychology*, 2015, https://apa.org/pubs/journals/features/xge-0000098.pdf.
56. Fazio et al, "Knowledge does not protect against illusory truth;" and Lynn Hasher and David Goldstein, "Frequency and the Conference of Referential Validity," *Journal of Verbal Learning and Verbal Behavior*, 1977, http://www.psych.utoronto.ca/users/hasherlab/PDF/Frequency%20and%20the%20conference%20Hasher%20et%20al%201977.pdf.
57. David M. J. Lazer et al, "The Science of Fake News," *Science*, 09 March 2018, http://science.sciencemag.org/content/359/6380/1094/tab-pdf.
58. Brendan Nyhan and Jason Reifler, "Misinformation and Fact-Checking: Research Findings from Social Science," New America Foundation, February 2012, https://www.dartmouth.edu/~nyhan/Misinformation_and_Fact-checking.pdf.
59. Drew B. Margolin, Aniko Hannak, and Ingmar Webber, "Political Fact-Checking on Twitter: When Do Corrections Have an Effect?," *Political Communication*, 2018, https://www.tandfonline.com/doi/full/10.1080/10584609.2017.1334018?scroll=top&needAccess=true; and Anisa Subedar, "The Godfather of Fake News," BBC, 27 November 2018, https://www.bbc.co.uk/news/resources/idt-sh/the_godfather_of_fake_news.
60. Alice Marwick, "Why Do People Share Fake News? A Sociotechnical Model of Media Effects," Georgetown Law Technology Review, 2018, https://www.georgetownlawtechreview.org/wp-content/uploads/2018/07/2.2-Marwick-pp-474-512.pdf.
61. Andrew Guess, Brendan Nyhan, and Jason Reifler, "Selective Exposure to Misinformation: Evidence from the consumption of fake news during the 2016 US presidential campaign," European Research Council, January 2018, https://www.dartmouth.edu/~nyhan/fake-news-2016.pdf.

One study of 7,200 participants found that counterfactual information can change partisan opinions when the information is presented with strong evidence.[62] Of note, this study also found that participants generally maintained an openness to opposing information unless they were primed to feel adversarial, or if the opposing arguments were laden with insulting language. Combining these insights with the aforementioned research on fight or flight reactions offers a potential roadmap for countering misleading information on polarizing topics: present corrective information in a tactful and well-supported manner that avoids disparaging those with opposing views.

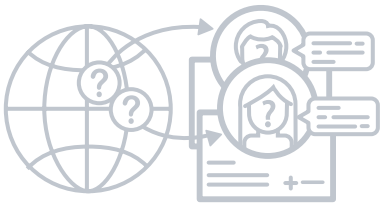
Research has also revealed different psychological interventions that may build resilience against disinformation. One study from the Cambridge Social Decision-Making Lab approached the topic of misinformation as a metaphorical "contagion." The study contended that misinformation spreads from person-to-person like a disease and therefore explored a potential immunization in the form of *attitudinal inoculation*.[63]

Similar to how a vaccine builds resistance to a virus, attitudinal inoculation warns people that they may be exposed to information that challenges their beliefs, before presenting a weakened example of the (mis)information and refuting it. This strategy can better inform, and even immunize, participants to similar misleading arguments in the future.[64] When applied to public attitudes about climate change, an experiment that used attitudinal inoculation with a polarized audience found that climate misinformation was less effective when participants were inoculated to similar misinformation in advance.[65]

Other research on cognitive ability examines *integrative complexity* (IC), which is a measure of a person's ability to accept and integrate multiple viewpoints. Low IC indicates a propensity for binary thinking and resistance to opposing perspectives, which has direct implications for the resonance of disinformation in polarized contexts.[66] To counter low IC, researchers have developed interventions to explore topics through the lens of different perspectives, which allows people to understand and overcome the cognitive biases that may render them adversarial toward opposing ideas. These interventions focus less on the *content* of one's thoughts and more on the *structure* of one's thoughts, therefore offering an approach that can be applied in many different contexts.[67]

As these combined strategies suggest, many of the same psychological factors that make humans susceptible to disinformation can also be used to defend against it. Repeating facts, offering solid evidence, preemptively warning about and debunking disinformation themes, and encouraging openness to differing viewpoints are all potential approaches for reducing vulnerabilities to disinformation.

62. Jin Woo Kim "Evidence Can Change Partisan Minds: Rethinking the Bounds of Motivated Reasoning," Job Market Paper, 30 September 2018, https://jinwookimqssdotcom.files.wordpress.com/2018/10/kim_ws.pdf.
63. Sander van der Linden et al, "Inoculating the Public against Misinformation about Climate Change," Wiley Online Library, 23 January 2017, https://onlinelibrary.wiley.com/doi/full/10.1002/gch2.201600008.
64. Van der Linden, "Inoculating the Public against Misinformation about Climate Change."
65. Van der Linden, "Inoculating the Public against Misinformation about Climate Change."
66. Peter Suedfeld, "The Cognitive Processing of Politics and Politicians: Archival Studies of Conceptual and Integrative Complexity," *Journal of Personality* 78, no. 6, October 2010, https://doi.org/10.1111/j.1467-6494.2010.00666.x.
67. Eolene M. Boyd-MacMillan et al., "I SEE! Scotland: Tackling Sectarianism and Promoting Community Psychosocial Health," *Journal of Strategic Security 9*, no. 4, 2016, http://dx.doi.org/10.5038/1944-0472.9.4.1556.



4. A look at foreign state-sponsored disinformation and propaganda

As the adoption of new technology and social media platforms have spread globally, so too have government efforts to exploit these platforms for their own interests, at home and abroad. Russian attempts to influence the United States 2016 presidential election and the 2016 Brexit vote in the United Kingdom are two recent, high-profile examples.

Yet the use of disinformation extends well beyond Russian interference in the US and the UK. A University of Oxford study found evidence of organized disinformation campaigns in 48 countries in 2018, up from 28 the year prior.[68]

Below is an overview of several countries notable for the extent and sophistication of their foreign influence and disinformation campaigns.

Russian influence and disinformation campaigns

Revelations of Russian interference in the lead-up to the 2016 US presidential election heightened the public's awareness of disinformation attacks against the United States. A 2017 report by the US Director of National Intelligence concluded that Russian President Vladimir Putin ordered an influence campaign that combined covert cyber operations (hacking, troll farms, and bots) with overt actions (dissemination of disinformation by Russian-backed media) in an effort to undermine public trust in the electoral process and influence perceptions of the candidates.[69]

The extent of this campaign was significant – thousands of Russian-backed human operatives and automated bots created more than one million tweets and hundreds of thousands of Facebook and

68. Samantha Bradshaw and Philip N. Howard, "Challenging Truth and Trust: A Global Inventory of Organized Social Media Manipulation," Oxford Internet Institute's Computational Propaganda Research Project, July 2018, http://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2018/07/ct2018.pdf.

69. US Office of the Director of National Intelligence, "Background to "Assessing Russian Activities and Intentions in Recent US Elections," The Analytic Process and Cyber Incident Attribution," 06 January 2017, https://www.dni.gov/files/documents/ICA_2017_01.pdf.

Instagram posts, while uploading more than 1,000 YouTube videos.[70] The tweets garnered 288 million views and the Facebook posts reached 126 million US accounts.[71]

Russia's efforts in 2016 may have come as a surprise to many, yet **Russian disinformation campaigns against the United States are not a recent phenomenon.**

Russia's efforts in 2016 may have come as a surprise to many, yet Russian disinformation campaigns against the United States are not a recent phenomenon. The Soviet Union deployed similar strategies called "active measures" during the Cold War era, which were coordinated efforts by Soviet organizations and intelligence agencies to influence opinions in foreign countries.[72] In the Soviet Union, propaganda was a key element of statecraft, as important as the work of the military or intelligence agencies.

During the late 1980s, for example, the Soviet Union coordinated a global disinformation campaign to convince the world's public that the United States had created the AIDS virus as a biological weapon.[73] This campaign intended to stir up anti-Americanism in the developing world and distract public attention from American charges that the Soviet Union was itself developing biological weapons.

The rumor, first published in 1983 via a Soviet-backed newspaper in India, appeared in Russian media in 1985 and was later published dozens of times in Russian media outlets. Broadcasts by Radio Moscow targeted at African countries claimed that America was deliberately spreading AIDS in Zaire, echoing previous claims by the outlet that the United States was developing biological weapons capable of killing only non-whites.

A similar campaign was mounted by the Soviets around the same time to spread unfounded allegations over child organ trafficking in the United States. The rumor was first reported by journalists during an interview of a Honduran government official in 1987.[74] Although the statement was quickly corrected by high-level Honduran officials, Russian media repeatedly ran the sensationalist rumors between 1987 and 1988 without mentioning the corrections. The allegations gained momentum over subsequent years, resulting in Turkey suspending its adoption program with the United States in the early 1990s. Not long after, a Guatemalan mob attacked US citizens in 1994 on the basis of this story.[75]

The rise of major social media platforms has offered the Kremlin an opportunity to innovate on this decades-old playbook. Under former President Dmitry Medvedev, the Russian government engaged in its

70. Ben Popkin, "Russian trolls went on attack during key election moments," NBC News, 20 December 2017, https://www.nbcnews.com/tech/social-media/russian-trolls-went-attack-during-key-election-moments-n827176; and: Mike Isaac and Daisuke Wakabayashi, "Russian influence reached 126 million Americans through Facebook alone," *New York Times*, 30 October 2017, https://www.nytimes.com/2017/10/30/technology/facebook-google-russia.html.
71. Isaac and Wakabayashi, "Russian influence reached 126 million Americans through Facebook alone."
72. Steve Abrams, "Beyond propaganda: Soviet active measures in Putin's Russia." Connections : *The Quarterly Journal*, 15(1), 31 May 2016, doi:http://dx.doi.org.ezproxy.cul.columbia.edu/10.11610/Connections.15.1-01
73. US Department of State, "Soviet Influence Activities: A Report on Active Measures and Propaganda, 1986-87," August 1987, https://www.globalsecurity.org/intell/library/reports/1987/soviet-influence-activities-1987.pdf.
74. US Information Agency, "The Child Organ Trafficking Rumor: A Modern 'Urban Legend'," A Report Submitted to the UN Special Rapporteur On The Sale Of Children, Child Prostitution, And Child Pornography, December 1994, http://pascalfroissart.online.fr/3-cache/1994-leventhal.pdf.
75. US Information Agency, "The Child Organ Trafficking Rumor: A Modern 'Urban Legend'," A Report Submitted to the UN Special Rapporteur On The Sale Of Children, Child Prostitution, And Child Pornography, December 1994, http://pascalfroissart.online.fr/3-cache/1994-leventhal.pdf.

Although many of the hallmarks of Soviet propaganda are present in Russia's modern-day propaganda efforts, **what has changed is the speed with which its narratives are created and disseminated.**

first widespread deployment of bots to spread political disinformation domestically.[76] This disinformation campaign proved difficult; two of Russia's more developed and competitive industries are technology and media, and the Russian blogosphere was not easily swayed by government disinformation. These challenges spurred the Russian government to develop highly-sophisticated automated bots and trolling techniques to better control the Russian public's conversations online. These same tools and their successors would later be unleashed on foreign adversaries.[77]

At present, 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 cyber criminals.[78] Although many of the hallmarks of Soviet propaganda are present in Russia's modern-day propaganda efforts, what has changed is the speed with which its narratives are created and disseminated.

Before 2016, Russia honed its online disinformation efforts in its immediate sphere of influence. As noted at the outset of this report, Russia deployed a coordinated online influence campaign during its annexation of Crimea in 2014. Russian state-controlled media outlets painted a uniquely anti-Ukrainian, pro-Russian narrative surrounding then-President Viktor Yanukovych's flight from Ukraine and the subsequent Russian invasion of Crimea.[79] To help shore up domestic support for Russia's actions, Russian government bots dominated the domestic political conversation in Russia during this period. Between 2014-2015, as much as 85 percent of the active Twitter accounts in Russia tweeting about politics were, in fact, government bots.[80]

In mid-2016, the Kremlin unleashed these tactics during the United Kingdom's successful June 2016 referendum vote to leave the European Union. One analysis of tweets found that in the 48 hours leading up to the vote, over 150,000 Russian accounts tweeted about #Brexit and posted more than 45,000 messages about the vote.[81] On the day of the referendum, Russian accounts tweeted 1,102 times with the hashtag #ReasonsToLeaveEU.[82] Meanwhile, Russia was deploying a similar strategy during the 2016 US presidential campaign.

The Kremlin-backed Internet Research Agency (IRA) initiated its efforts to interfere in US politics as early as 2014, spending $1.25 million per month on its combined domestic and global operations, which included

76. Sergey Sanovich, "Computational Propaganda in Russia: The Origins of Digital Disinformation," Eds: Samuel Woolley and Philip N. Howard, Working Paper [Oxford, UK: Project on Computational Propaganda, March 2017] http://comprop.oii.ox.ac.uk/.
77. Sanovich, "Computational Propaganda in Russia: The Origins of Digital Misinformation."
78. Alina Polyakova and Spencer P. 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/wp-content/uploads/2018/03/the-future-of-political-warfare.pdf.
79. Tomila Lankina and Kohei Watanabe, "Russian Spring' or 'Spring Betrayal'? The Media as a Mirror of Putin's Evolving Strategy in Ukraine," *Europe-Asia Studies*, March 2018, https://doi.org/10.1080/09668136.2017.1397603.
80. Denis Stukal et al, "Detecting Bots on Russian Political Twitter," *Big Data*, December 2017, https://www.liebertpub.com/doi/10.1089/big.2017.0038.
81. UK Parliament, "Russian influence in political campaigns," Disinformation and 'fake news': Interim Report, https://publications.parliament.uk/pa/cm201719/cmselect/cmcumeds/363/36308.htm.
82. Matthew Field and Mike Wright, "Russian trolls sent thousands of pro-Leave messages on day of Brexit referendum, Twitter data reveals, *The Telegraph*, 17 October 2018, https://www.telegraph.co.uk/technology/2018/10/17/russian-iranian-twitter-trolls-sent-10-million-tweets-fake-news/.

During the 2016 presidential campaign, **Russian posts reached 126 million US Facebook accounts**.

dedicated English language staff focused on the 2016 US presidential campaign.[83] The secretive agency was headquartered in a heavily-guarded building in downtown St. Petersburg.[84] On one floor, employees produced a high volume of fake articles, using mostly original text to create a veneer of authenticity, and on another floor a separate group of employees created fake social media accounts to distribute these articles and then post comments about them.[85]

An NBC report identified 2,752 Russian "troll" accounts that posted more than 200,000 tweets; these tweets earned 2.1 million retweets and 1.9 million likes.[86] Twitter reported an even more expansive campaign that likely extended beyond the IRA, with 36,000 automated accounts posting 1.4 million tweets that earned 288 million views leading up to the election.[87] On Facebook, Russian posts reached 126 million US Facebook accounts. On Instagram, which is wholly owned by Facebook, 170 Russian accounts created more than 120,000 pieces of content, which reached more than 20 million US accounts.[88] The activities of the IRA were not limited to Facebook, Instagram, and Twitter; it also targeted YouTube, Google+, Vine, Meetup, Pinterest, Tumblr, Gab, Medium, Reddit, and even PayPal, which helped sell its merchandise.[89]

The IRA's activities on Instagram were particularly effective at generating impressions. Instagram's platform is conducive for posting the most viral content – jokes and memes – and Russian accounts leveraged this platform to maximize their reach. Between 2014 and 2017, IRA content on Instagram reached 187 million engagements (likes and shares), far exceeding their content's 76.5 million engagements on Facebook.[90] The New Knowledge Report on the Internet Research Agency's disinformation tactics predicts that "Instagram is likely to be a key battleground on an ongoing basis."[91]

It is clear that there was a significant volume of Russian posts and impressions generated during the 2016 US presidential campaign. However, some have cautioned against exaggerating the impact of Russian disinformation on the outcome of the election.[92]

83. US Department of Justice, "United States of America vs. Internet Research Agency," filed 16 February 2018, https://www.justice.gov/file/1035477/download.
84. Ben Popken and Kelly Cobiella, "Russian troll describes work in the infamous misinformation factory," *NBC News*, 16 November 2017, https://www.nbcnews.com/news/all/russian-troll-describes-work-infamous-misinformation-factory-n821486.
85. Popken and Cobiella, "Russian troll describes work in the infamous misinformation factory."
86. Ben Popken, "Russian trolls went on attack during key election moments" *NBC News*, 20 December 2017, https://www.nbcnews.com/tech/social-media/russian-trolls-went-attack-during-key-election-moments-n827176.
87. Mike Isaac and Daisuke Wakabayashi, "Russian influence reached 126 million Americans through Facebook alone," *New York Times*, 30 October 2017, https://www.nytimes.com/2017/10/30/technology/facebook-google-russia.html.
88. Renee DiResta et al, "The Tactics and Tropes of the Internet Research Agency," New Knowledge, December 2018, https://disinformationreport.blob.core.windows.net/disinformation-report/NewKnowledge-Disinformation-Report-Whitepaper.pdf; and Isaac and Wakabayashi, "Russian influence reached 126 million Americans through Facebook alone."
89. DiResta et al, "The Tactics and Tropes of the Internet Research Agency," December 2018; and Philip N. Howard et al, "The IRA, Social Media and Political Polarization in the United States, 2012-2018" [Oxford, UK: Project on Computational Propaganda, 2018], https://comprop.oii.ox.ac.uk/wp-content/uploads/sites/93/2018/12/IRA-Report-2018.pdf.
90. DiResta et al, "The Tactics and Tropes of the Internet Research Agency."
91. DiResta et al, "The Tactics and Tropes of the Internet Research Agency."
92. Patrick Ruffini, "Why Russia's Facebook ad campaign wasn't such a success," *Washington Post*, 3 November 2017, https://www.washingtonpost.com/outlook/why-russias-facebook-ad-campaign-wasnt-such-a-success/2017/11/03/b8efacca-bffa-11e7-8444-a0d4f04b89eb_story.html?utm_term=.78cb0da3de03.

Most Americans, for example, probably only scrolled past a very small number of Russian-backed posts throughout the duration of the campaign, which says nothing about whether they read, clicked on, or were influenced in any meaningful way by the content. Furthermore, the several hundred million impressions of Russian propaganda across Twitter and Facebook during the campaign were dwarfed by the billions of total *daily* impressions of all content across both platforms. Kremlin-generated impressions were a drop in the bucket compared to total user activity, which calls into question their ability to have played a decisive role in swaying public opinion.

In the wake of the 2016 election, the **Kremlin appears intent on continuing to leverage disinformation to influence political discourse** in the United States and elsewhere.

Russia's ad-targeting also appeared to lack an overarching electoral strategy. For example, less than $2,000 was spent on Russian ads in the battleground state of Wisconsin, and even less on the battleground states of Pennsylvania and Michigan, suggesting that Russian content did not deliver meaningful impact on the electoral college votes that decided the election.[93] Others have argued that the IRA's disinformation campaign was amateurish and careless, even failing to hide the origin of its content, which further underscores the need for caution when assessing the effectiveness of its propaganda.[94]

It is perhaps more plausible that Russian cyberhacks into the Clinton campaign - rather than the Kremlin's social media disinformation - impacted the course of the election. Kathleen Jamieson, the director of the Annenberg Public Policy Center at the University of Pennsylvania, has argued that the disclosures from WikiLeaks' release of Russian-hacked Clinton campaign emails caused the decline in polled voters' trust in Clinton in October 2016.[95]

In the wake of the 2016 election, the Kremlin appears intent on continuing to leverage disinformation to influence political discourse in the United States and elsewhere. Indeed, US sanctions and condemnations seem to have done little to dissuade the Russians from maintaining these efforts. While the IRA spent $12 million during the 2016 election campaign, its budget totaled $12.2 million in 2017 and $10 million for the first half of 2018 leading up the US midterms.[96] Russian posters have also adapted their tactics, shifting away from producing fictional content which can be censored by platform moderators, towards amplifying existing political memes promoted by far-right and far-left sources.[97]

Russia-backed Syrian influence and disinformation campaigns

Pro-Syrian government disinformation has proliferated through social media since the Syrian conflict began in 2011. Much of the disinformation appears to be produced and disseminated by Kremlin-linked

93. Alexis C. Madrigal, "Russia's Troll Operation Was Not That Sophisticated," *The Atlantic*, 19 February 2018, https://www.theatlantic.com/technology/archive/2018/02/the-russian-conspiracy-to-commit-audience-development/553685/.
94. Neil MacFarquhar, "Russian Trolls Were Sloppy, but Indictment Still 'Points at the Kremlin'," *New York Times*, 17 February 2018, https://www.nytimes.com/2018/02/17/world/europe/russia-indictment-trolls-putin.html.
95. Jane Mayer, "How Russia Helped Swing the Election for Trump," ***The New Yorker***, 1 October 2018, https://www.newyorker.com/magazine/2018/10/01/how-russia-helped-to-swing-the-election-for-trump.
96. US Department of Justice, "United States of America v. Elena Alekseevna Khusyaynova," filed 28 September 2018, https://www.justice.gov/usao-edva/press-release/file/1102591/download.
97. Joseph Menn, "Russia seen adopting new tactics in US election interference efforts," ***Reuters***, 06 November 2018, https://www.reuters.com/article/us-usa-election-russia/russia-seen-adopting-new-tactics-in-u-s-election-interference-efforts-idUSKCN1NB0PI?feedType=RSS&feedName=worldNews.

accounts as an extension of Russia's military support for the Assad regime. The following are a few notable examples of the disinformation tied to the ongoing conflict.

In April 2018, in the wake of a sarin gas attack in Idlib Province, there was a wave of alt-right online postings in the United States claiming the attack was a "false flag" operation intended to undermine the Assad regime. The Atlantic Council traced these alt-right postings back to a pro-Assad outlet, *Al-Masdar News*, which had first published the allegation in the attack's immediate aftermath.[98] In retaliation for the sarin gas attack, the United States launched a strike against the Syrian regime. The Pentagon reported an observed 2,000 percent increase in Russian bot activity spreading disinformation on Twitter in the 24 hours after the US strike.[99]

Several months later, this pattern of online media disinformation repeated. In September 2018, Assad's regime announced its intention to launch a military assault on the rebel-held Idlib province. As Trump warned Syria against an anticipated chemical weapons attack, Twitter saw a surge in fake Russian account creation, ostensibly preparing to spread disinformation around a future attack.[100] Concurrently, the Russian Ministry of Defense released a statement claiming the Syrian rebels were preparing a false flag operation to gas their own people and have it filmed by the White Helmets, a volunteer organization created to conduct search and rescue operations in rebel-held Syria.[101] The Ministry of Defense's allegations were later amplified by *Russia Today*, the Kremlin's external media arm.[102]

A 2018 inquiry by the Bellingcat investigation team found that the White Helmets became a target of a "brutal and unrelenting" disinformation campaign because they provided evidence of gross human rights violations by both the Syrian government and the Russian military.[103] The campaign, which was planned and unleashed by the Russian and Syrian governments, portrayed the White Helmets as terrorists, agents of the West, and "legitimate targets."[104]

In support of their propaganda campaign, Russian agencies and media outlets published numerous reports starting in February 2018 that made unsubstantiated claims about the White Helmets transporting chemical weapons to Idlib province.[105] In December 2018, the US Department of State issued a statement

98 Digital Forensic Research Lab, "How the alt-right brought #SyriaHoax to America," Medium Blog, 07 April 2017, https://medium.com/dfrlab/how-the-alt-right-brought-syriahoax-to-america-47745118d1c9.

99. Jessica Kwong, "Russian Trolls Increased '2,000 Percent' After Syria Attack, Pentagon Says," *Newsweek*, 14 April 2018, https://www.newsweek.com/russian-trolls-increased-2000-percent-after-syria-attack-pentagon-says-886248.

100. Jack Nassetta and Ethan Fecht, "Russia is gearing up to misinform the US public about Syria. Here's our cheat sheet to identify Twitter trolls," *Washington Post*, 17 September 2018, https://www.washingtonpost.com/news/monkey-cage/wp/2018/09/17/russia-is-gearing-up-to-misinform-the-u-s-public-about-syria-heres-our-cheat-sheet-to-identify-twitter-trolls/?noredirect=on&utm_term=.24ff76b55894.

101. Josie Ensor, "Russian misinformation about 'imminent' White Helmets chemical attack could spell start of Idlib seige," *Telegraph*, 02 September 2018, https://www.telegraph.co.uk/news/2018/09/02/russian-disinformation-campaign-syria-threatened-spark-new-war/.

102. "Filming of staged chemical attack in Idlib, Syria begins – Russian MoD," *RT News*, 11 September 2018, https://www.rt.com/news/438158-staged-chemical-attack-idlib/.

103. Bellingcat Investigation Team, "Chemical Weapons and Absurdity: The Disinformation Campaign Against the White Helmets," A Joint Report in Collaboration with Newsy, 18 December 2018, https://www.bellingcat.com/news/mena/2018/12/18/chemical-weapons-and-absurdity-the-disinformation-campaign-against-the-white-helmets/.

104. Bellingcat Investigation Team, "Chemical Weapons and Absurdity."

105. Louisa Loveluck, "Russian disinformation campaign targets Syria's beleaguered rescue workers," *Washington Post*, 18 December 2018, https://www.washingtonpost.com/world/russian-disinformation-campaign-targets-syrias-beleaguered-rescue-workers/2018/12/18/113b03c4-02a9-11e9-8186-4ec26a485713_story.html?utm_term=.8733e7fd7096.

that a "chemical" attack in Aleppo reported by Syrian and Russian state media was actually a tear-gas attack conducted by Syrian government forces to "undermine confidence in the ceasefire in Idlib."[106]

Chinese influence and disinformation campaigns

In September 2018, a four-page advertisement sponsored by the state-owned *China Daily* ran in the *Des Moines Register*. The advertisement, mirroring an actual newspaper spread with journalistic articles, included a selection of pieces that touted the benefits of free trade for US farmers, the economic risks of China-US trade tensions, and President Xi's long ties to the state of Iowa.[107] Targeting an Iowa audience in the midst of China's agricultural trade dispute with Trump – and during the midterm campaign season no less – made clear that China would not hesitate to try shaping the US political conversation.

As the *Des Moines Register* example illustrates, China's propaganda efforts are distinct from Russia's in several respects. While Russia's campaigns tend to be focused on the cyber domain, as evidenced by its 2016 election interference efforts, China's international influence campaigns are largely characterized by economic, political, and personal relationship-building.[108] Chinese campaigns have been widespread. They range from the production and global distribution of pro-Chinese media, to attempts to influence educational and policy institutions abroad, to the wielding of financial influence through aggressive loans and infrastructure investment.[109]

In 2010, China launched a television, radio, and print media campaign to change how consumers of news viewed the country and its place in the world. This included nearly tripling the number of China Central Television (CCTV) bureaus globally, increasing the number of foreign correspondents for the *China Daily*, and building out an English-language tabloid, the *Global Times*.[110] China's English-language outlets have produced hundreds of articles touting China's prosperity and stability, aimed primarily at a foreign audience.[111] In 2018, President Xi announced the merger of CCTV, China Radio International, and China National Radio into a single network named Voice of China. The new network's mission includes strengthening international outreach and influencing public opinion abroad.[112]

106. Bureau of Public Affairs, US Department of State, "The Russian and Assad Regime's False Allegations on Chemical Weapons Use in Aleppo," 07 December 2018, https://www.state.gov/r/pa/prs/ps/2018/12/287941.htm.
107. Donnelle Eller, "Chinese-backed newspaper insert tries to undermine Iowa farm support for Trump, trade war," *Des Moines Register*, 24 September 2018, https://www.desmoinesregister.com/story/money/agriculture/2018/09/24/china-daily-watch-advertisement-tries-sway-iowa-farm-support-trump-trade-war-tariffs/1412954002/.
108. Abigail Grace, "China's Influence Operations Are Pinpointing America's Weaknesses," *Foreign Policy*, 04 October 2018, https://foreignpolicy.com/2018/10/04/chinas-influence-operations-are-pinpointing-americas-weaknesses/; and Andrea Kendall-Taylor and David Shullman, "How Russia and China Undermine Democracy," *Foreign Affairs*, 02 October 2018, https://www.foreignaffairs.com/articles/china/2018-10-02/how-russia-and-china-undermine-democracy.
109. Samantha Custer et al, "Ties That Bind: Quantifying China's public diplomacy and its "good neighbor" effect," [Williamsburg, VA: AidData at William & Mary, 2018] http://docs.aiddata.org/ad4/pdfs/Ties_that_Bind—Executive_Summary.pdf.
110. "The Chinese Are Coming," *The Economist*, 04 March 2010, https://www.economist.com/asia/2010/03/04/the-chinese-are-coming.
111. Paul Mazur, "China Spreads Propaganda to US on Facebook, a Platform It Bans at Home," *New York Times*, 08 November 2017, https://www.nytimes.com/2017/11/08/technology/china-facebook.html?mtrref=www.google.com&mtrref=undefined&gwh=16B98CC8D9E3ABFA00633192121284B5&gwt=pay.
112. "China state media merger to create propaganda giant," *The Guardian*, 21 March 2018, https://www.theguardian.com/world/2018/mar/21/china-state-media-merger-to-create-propaganda-giant

Through China's "United Work Front Department," the Chinese Communist Party (CCP) has attempted to influence leading academic institutions and think tanks.[113] Delegations of Chinese authorities focused on Tibet have paid visits to universities with influential Tibetan academics, including Columbia, Harvard, and the University of Virginia, to exchange views and share official CCP talking points. One such example included Chinese consular officials in New York City twice paying visits to Columbia Tibetan professor Robert Barnett, threatening to cut off future communication if he did not better align his views on Tibet with the CCP.[114]

The CCP-linked China-United States Exchange Foundation (CUSEF) has partnered with, and in some cases funded, institutions including the Brookings Institution, the Center for Strategic and International Studies, the Atlantic Council, the East-West Institute, the Carter Center, the Carnegie Endowment, and Johns Hopkins University.[115] In addition, Chinese educational centers and events have dramatically expanded in recent years. Confucius Institutes, located on university campuses abroad, have grown to more than 500 globally and have been used to apply pressure on professors and censor what Chinese professors can teach abroad.[116] Certain universities have pushed back, including the University of Chicago, which removed its Confucius Institute from campus in 2014 after 100 professors signed a petition protesting its presence.[117]

China's efforts to spread influence extends to social media platforms that have been banned within its own borders. China's domestic internet controls have long been robust, coupling censorship of sensitive topics with an outright ban of many western social media and technology platforms (among them Google, Facebook, and YouTube). This "Great Firewall" has continued to grow in 2018, with the government recently expanding cybersecurity laws and advancing its surveillance capabilities, while making it increasingly difficult to use virtual private network services to avoid the firewall.[118] While Chinese social media is allowed, it is both tightly controlled by government censors and flooded with propaganda. Some two million government-paid individuals contribute roughly 450 million pro-government posts annually to distract and drown out any domestic criticisms of the CCP.[119]

China's ban on western social media is a tacit acknowledgment of these platforms' potential to influence Chinese citizens. Meanwhile, the CCP uses foreign platforms' networks to spread state-sponsored advertisements in foreign countries, including the United States. Chinese entities are Facebook's largest ad-buyers in Asia, even though Chinese citizens cannot use the platform.[120] Some estimate Chinese buyers

113. Alexander Bowe, "China's Overseas United Front Work: Background and Implications for the United States," US-China Economic and Security Review Commission, 24 August 2018, https://www.uscc.gov/sites/default/files/Research/China%27s%20Overseas%20United%20Front%20Work%20-%20Background%20and%20Implications%20for%20US_final_0.pdf.
114. Anastasya Lloyd-Damnjanovic, "A Preliminary Study of PRC Political Influence and Interference Activities in American Higher Education," Wilson Center, August 2018, https://www.wilsoncenter.org/sites/default/files/prc_political_influence_full_report.pdf.
115. Bowe, "China's Overseas United Front Work: Background and Implications for the United States."
116. Rachelle Peterson, "American Universities Are Welcoming China's Trojan Horse," *Foreign Policy*, 09 May 2017, https://foreignpolicy.com/2017/05/09/american-universities-are-welcoming-chinas-trojan-horse-confucius-institutes/.
117. Te-Ping Chen, "Thanks, But No Thanks, University of Chicago Tells Confucius Institute," *Wall Street Journal*, 26 September 2014, https://blogs.wsj.com/chinarealtime/2014/09/26/thanks-but-no-thanks-university-of-chicago-tells-confucius-institute/.
118. Adrian Shahbaz, "Fake news, data collection, and the challenge to democracy" in *Freedom on the Net 2018: The Rise of Digital Authoritarianism*, Freedom House, 2018, https://freedomhouse.org/report/freedom-net/freedom-net-2018/rise-digital-authoritarianism.
119. Gary King, Jennifer Pan, and Margaret E. Roberts, "How the Chinese Government Fabricates Social Media Posts for Strategic Distraction, not Engaged Argument, *American Political Science Review*, 2017, pp. 484-501, https://gking.harvard.edu/50C.
120. Mazur, "China Spreads Propaganda to US on Facebook, a Platform It Bans at Home."

spent five billion dollars on Facebook ads in 2018, making them the second largest market after the United States.[121] State-sponsored media ads, a small fraction of that total, mirror the CCP's offline efforts to paint positive portrayals of the Chinese government and broader Chinese society. Furthermore, the CCP has established Facebook pages for its various state-run media outlets, where content highlighting Chinese successes is distributed and promoted through page followers and paid advertising.

The CCP uses foreign platforms' networks to spread state-sponsored advertisements in foreign countries, including the United States.

Increasingly, China has interfered in foreign states in a manner similar to Russia. One of the more recent and aggressive examples is the November 2018 election campaign in Taiwan. The Chinese government undertook a substantial propaganda effort to undermine Taiwanese President Tsai Ing-wen and her Democratic Progressive Party (DPP) in hopes of ousting them from office in favor of the Kuomintang opposition party, who are seen as more compliant to Beijing's will.[122]

The Chinese engaged in a months-long campaign of anti-Tsai, anti-DPP social media messaging using platforms including Facebook, Twitter, and chat groups in a manner similar to Russia's influence campaigns.[123] Although difficult to ascertain the full impact of this campaign, Tsai's party did suffer a significant defeat during the election, prompting her to step down as leader of the DPP and improving Kuomintang's chances to retake the government.

Covert social media influence tactics have also been used by the CCP. Chinese spies have created fake LinkedIn accounts to connect and build relationships with western business leaders and academics. After organizing in-person meetings under false pretenses, these Chinese officials have made financial offers in exchange for establishing intelligence-sharing relationships.[124]

China's influence campaign is vast and multifaceted, but its ability to change minds has been questioned. Much of Chinese media content is so clearly biased in favor of the CCP that international audiences may have little trouble understanding these outlets' true intentions and are therefore unlikely to be swayed in any meaningful way.[125]

Similarly, a lack of objectivity on the part of Chinese media employees, such as the 2018 example of a China Global Television Network reporter verbally and physically harassing attendees at a UK press conference for being "traitors" and "fake Chinese," has led some to conclude that Chinese journalists' efforts are aimed more at impressing their superiors than truly swaying hearts and minds overseas.[126]

121. Preeti Varathan, "China, where Facebook is banned, could make $5 billion for the company this year," Quartz, 16 May 2018, https://qz.com/1279700/china-is-facebooks-second-largest-ad-spender-after-the-us/.
122. Chris Horton, "Specter of Meddling by Beijing Looms Over Taiwan's Elections," *New York Times*, 22 November 2018, https://www.nytimes.com/2018/11/22/world/asia/taiwan-elections-meddling.html.
123. Josh Rogin, "China's interference in the 2018 elections succeeded — in Taiwan," *Washington Post*, 18 December 2018, https://www.washingtonpost.com/opinions/2018/12/18/chinas-interference-elections-succeeded-taiwan/?utm_term=.f0bf442eab64.
124. Jonas Parello-Plesner, "China's LinkedIn Honey Traps," *The American Interest*, 23 October 2018, https://www.the-american-interest.com/2018/10/23/chinas-linkedin-honey-traps/.
125. Hilton Yip, "China's $6 Billion Propaganda Blitz is a Snooze," *Foreign Policy*, 23 April 2018, https://foreignpolicy.com/2018/04/23/the-voice-of-china-will-be-a-squeak/.
126. James Palmer, "China's Global Propaganda Is Aimed at Bosses, Not Foreigners," *Foreign Policy*, 01 October 2018, https://foreignpolicy.com/2018/10/01/chinas-global-propaganda-is-aimed-at-bosses-not-foreigners/.

Nonetheless, a significant portion of Chinese-language media outlets abroad have direct or indirect ties to members of the CCP, and consequently these outlets pump out a disproportionately large amount of pro-China content. This leaves international Chinese-language communities particularly exposed to this information.[127]

Iranian influence and disinformation campaigns

Between August and October 2018, Facebook cracked down on two separate Iranian propaganda campaigns, removing hundreds of Facebook and Instagram accounts, pages, and groups, some of which dated back to 2011. The pages alone were followed by more than one million Facebook accounts.[128]

Following the August crackdown, Twitter in turn removed hundreds of accounts that had been engaged in "coordinated manipulation."[129] On Reddit, users began noticing a pattern of suspicious posts beginning in July 2017 which targeted the *WorldNews* subreddit's 19 million followers. The posts included political content linked to obscure websites that Reddit moderators were able to trace back to Iran after investigating.[130] The cybersecurity company FireEye released its own report that concluded Iran was behind a coordinated disinformation campaign, targeting US and UK audiences and dating back to at least 2017.[131]

Iranian disinformation has displayed similarities to both Russian and Chinese tactics. For example, they leveraged fake Twitter, Facebook, Instagram, and Reddit accounts to coordinate disinformation during the 2018 midterms in a manner similar to the Kremlin's efforts during the 2016 and 2018 US election campaigns.

Like China, Iranian propaganda efforts have largely focused on promoting particular policy interests that are aligned with the Iranian government. Iranian interests promoted by these accounts include anti-Israeli and pro-Palestinian narratives, as well as condemnations of Iran's adversary Saudi Arabia. Likewise, there has been an overwhelming anti-Trump bias to their content, which some have perceived as a response to President Trump's hardline rhetoric towards Iran and his decision to withdraw the United States from the 2015 nuclear agreement with Iran.[132]

A Reuters investigation in November 2018 found that Iran used more than 70 propaganda websites to spread disinformation to 15 countries, including the US and UK. These websites, which had more than

127. Hilton Yip, "China's $6 Billion Propaganda Blitz is a Snooze."

128. "Taking Down Coordinated Inauthentic Behavior from Iran," *Facebook Newsroom*, 26 October 2018, https://newsroom.fb.com/news/2018/10/coordinated-inauthentic-behavior-takedown/; and Craig Timberg et al, "Sprawling Iranian influence operation globalizes tech's war on disinformation," *Washington Post*, 21 August 2018, https://www.washingtonpost.com/technology/2018/08/21/russian-iran-created-facebook-pages-groups-accounts-mislead-users-around-world-company-says/?noredirect=on&utm_term=.aa3fbb707c54.

129. Timberg, et al, "Iranian influence operation globalizes tech's war on disinformation."

130. Ben Collins, "Volunteers found Iran's propaganda effort on Reddit — but their warnings were ignored," NBC News, 24 August 2018, https://www.nbcnews.com/tech/tech-news/volunteers-found-iran-s-propaganda-effort-reddit-their-warnings-were-n903486.

131 "Suspected Iranian Influence Operation: Leveraging Inauthentic News Sites and Social Media Aimed at US, UK, Other Audiences," Fireeye, 21 August 2018, https://www.fireeye.com/content/dam/fireeye-www/current-threats/pdfs/rpt-FireEye-Iranian-IO.pdf.

132. Collins, "Volunteers found Iran's propaganda effort on Reddit — but their warnings were ignored;" and Issie Lapowsky, "What We Now Know About Iran's Global Propaganda Campaign," *Wired*, 24 August 2018, https://www.wired.com/story/iran-global-propaganda-fireeye/.

One Iranian news site concocted a fake piece in late 2016 alleging that the Israeli government threatened a nuclear attack if Pakistan sent troops to Syria. **Failing to realize the story was fake, Pakistan's then-Defense Minister responded with an *actual* nuclear threat against Israel** via Twitter.

500,000 monthly visitors, were promoted on social media by Iranian accounts with over one million followers.[133]

One such site, AWDnews, concocted a fake piece in late 2016 alleging that the Israeli government threatened a nuclear attack if Pakistan sent troops to Syria. Failing to realize the story was fake, Pakistan's then-Defense Minister responded with an *actual* nuclear threat against Israel via Twitter. Cooler heads prevailed once the initial piece was revealed to be a hoax.[134]

Iranian pages have also found success in generating followers by doctoring memes, often around polarizing topics. One popular page with more than 400,000 likes, titled "No racism no war," photoshopped an image of Tom Hanks by adding a Black Lives Matter slogan to his t-shirt. The image generated 95,000 shares.[135] Since 2011, well-produced, fake BBC Persian videos have been created to cover stories and provide analysis inconsistent with BBC Persian's actual content. These videos are posted to websites that are prominent in search results for BBC Persian. The videos have also been spread through pro-Iranian social media pages.[136]

Military propaganda plays a prominent role in Iranian influence operations as well. Iran's Ministry of Intelligence and National Security, which coordinates this disinformation, has released reports exaggerating Iranian military strength and technological developments, hoping to obscure and complicate accurate assessments of their capabilities. By bolstering the perception of Iranian military might overseas, Iran believes it can weather international pressure and help deter future military threats. However, like China, some analysts view Iran's overt efforts at propaganda to be too conspicuous to effectively sway international audiences.[137]

North Korean influence and disinformation campaigns

The North Korean government (DPRK) has engaged in disinformation efforts to not only influence international actors and spread pro-DPRK propaganda, but also to skirt international sanctions. One example of the former was North Korea's hacking attack against Sony Pictures in 2014. It was a brazen attempt to blackmail the company into cancelling the release of the parody film *The Interview*, and it illustrated the DPRK's willingness to try to silence critics through criminal cyber activity.[138]

133. Jack Stubbs and Christopher Bing, "Special Report: How Iran Spreads Disinformation Around the World," *Reuters, 30 November 2018*, https://www.reuters.com/article/us-cyber-iran-specialreport-idUSKCN1NZ1FT.
134. "Iran duped Pakistan into Israel nuke threat as tiny part of huge fakery campaign," *The Times of Israel*, 30 November 2018, https://www.timesofisrael.com/iran-duped-pakistan-into-israel-nuke-threat-as-tiny-part-of-huge-fakery-campaign/.
135. Alexis C. Madrigal, "Iranian Propaganda Targeted Americans With Tom Hanks," *The Atlantic*, 26 October 2018, https://www.theatlantic.com/technology/archive/2018/10/irans-facebook-propaganda-targeted-americans-tom-hanks/574129/.
136. Thomas Brewster, "Inside The 7-Year-Old Iranian Propaganda Machine Producing Fake BBC News," *Forbes*, 28 February 2018, https://www.forbes.com/sites/thomasbrewster/2018/02/28/bbc-iran-fake-news/#3ec2371b54f1.
137. Firas Elias, "Iran's Military Propaganda: Failures and Successes," The Washington Institute for Near East Policy, 10 September 2018, https://www.washingtoninstitute.org/fikraforum/view/irans-military-propaganda-failures-and-successes.
138. Tim Starks, "US indicts North Korean national for Sony hack, massive cyberattacks," *Politico*, 6 September 2018, https://www.politico.com/story/2018/09/06/justice-department-north-korea-sony-hack-771212.

As of 2018, **an estimated several hundred DPRK agents operate fake cyber accounts to influence online discourse** in favor of the regime.

Covert tactics have also been used extensively by North Korea. As of 2018, an estimated several hundred DPRK agents operate fake cyber accounts to influence online discourse in favor of the regime. Aimed at a South Korean audience, these agents have created accounts meant to appear South Korean in order to post pro-DPRK comments, blog posts, and videos. Pyongyang's intent is twofold: to paint North Korea in a favorable light and to stoke division in South Korea.[139]

The DPRK has further leveraged covert disinformation tactics for nontraditional means - to fund the regime in the face of international sanctions pressure. North Koreans based in places like China have created a web of fake profiles and businesses on professional networking and freelancing platforms to deceive viewers and earn IT contracting business from clients around the globe. This money is sent back to the North Korean regime at a time when the DPRK is desperate for funds. By misrepresenting their identities, these illicit government-backed businesses have skirted international sanctions and funneled potentially millions of dollars to the DPRK.[140]

139. Tae-jun Kang, "North Korea's Influence Operations, Revealed," *The Diplomat*, 25 July 2018, https://thediplomat.com/2018/07/north-koreas-influence-operations-revealed/.

140. Wenxin Fan, Tom Wright, and Alastair Gale, "Tech's New Problem: North Korea," *Wall Street Journal*, 14 September 2018, https://www.wsj.com/articles/north-koreans-exploit-social-medias-vulnerabilities-to-dodge-sanctions-1536944018.

5. Platform-specific challenges and efforts to counter disinformation

Having examined how foreign states are weaponizing disinformation, this review now turns to the major technology platforms where disinformation and propaganda are disseminated. What are companies like Google, Facebook and Twitter doing to counter the problem?

This is an essential question for the simple reason that to neutralize online disinformation, the platforms themselves must play a central role. Their ability to solve this problem, at the moment, far exceeds that of any other organization – including national governments.

Companies like Google, Facebook and Twitter possess the overwhelming majority of the data pertaining to this issue. They have developed the proprietary algorithms that identify how, when, and where information is viewed by users. They have the most experience working on these issues and are best-equipped to improve how disinformation is tracked and countered. And they possess unparalleled levels of technical and monetary resources to address these issues; as of February 2019, Google, Facebook, and Twitter have a combined market valuation well above one trillion dollars.

Tech giants have demonstrated, to some extent, that they are willing to address disinformation. Google, Facebook, and Twitter are each building their own internal and external teams of security personnel and fact-checkers to counter propaganda and illegal content. They are also developing proprietary tools, some using artificial intelligence (AI) and machine learning, to limit disinformation on their platforms. Facebook, for instance, announced plans in 2018 to open two new AI labs that will focus, in part, on how to counter disinformation.[141]

Yet, social media platforms' incentives are not always prioritized to limit disinformation. In some respects, their incentives are aligned with spreading more of it. Tech giants' revenues are generated almost entirely through advertising, which depends on maximizing user engagement with the platform. As outlined earlier, users are more likely to click on or share sensational and inaccurate content; increasing clicks and shares translates into greater advertising revenue. The short-term incentives, therefore, are for the platforms to increase, rather than decrease, the amount of disinformation their users see.

141. Cade Metz, "Facebook Adds A.I. Labs in Seattle and Pittsburgh, Pressuring Local Universities," *New York Times*, 04 May 2018, https://www.nytimes.com/2018/05/04/technology/facebook-artificial-intelligence-researchers.html.

In the long term, however, the spread of disinformation and the growing public outcries against it may decrease public trust in social media brands and trigger more decisive action by these companies. As outlined below, the heightened public awareness and scrutiny leveled at social media platforms following the 2016 US presidential election spurred them to enact and overhaul a number of policies designed to counter online disinformation.

This section details how the major social media platforms have responded to disinformation and outlines some of the challenges that remain.

Although this report attempts to capture the most recent efforts of each platform, these companies frequently update user policies, algorithms, and strategies for countering disinformation. This assessment, therefore, should be viewed as a representative snapshot of the platform landscape in late 2018 and early 2019.

Countering disinformation at Facebook

As Russian and Chinese propaganda efforts have made clear, Facebook is vulnerable to the spread of disinformation and influence through both covert and overt strategies. Bots and trolls can create fake accounts to spread fake text, image, and video content through posts, pages, and paid advertising. At the same time, state-backed media companies often spread overt influence through legitimate accounts and pages that use paid advertising to disseminate their messages.

In the wake of the revelation that disinformation was shared widely over Facebook during the 2016 presidential campaign, Facebook announced the rollout of several features to help combat disinformation.[142] First, the company made it easier to flag false news to platform administrators by allowing users to click on the news post in question and select from pre-set reporting options.

Facebook also began enlisting the help of third-party fact-checkers to review reports from the community. After reviewing, these third-party fact-checkers provide a rating on the trustworthiness of an article. News deemed inaccurate automatically appears lower on a user's newsfeed and previously displayed a "disputed" label to warn users who may read or share the story on Facebook.[143] Facebook additionally began testing changes to its newsfeed rankings, weighing its algorithm against articles that present disinformation warning signs.

Concurrently, Facebook began reducing the financial incentives for false news. That included preventing fake news sites from "spoofing" the domains of real ones. A malicious actor can no longer imitate a legitimate news site's domain in an attempt to deceive readers.

Facebook's actions were not limited to the immediate wake of the 2016 race. In 2018, they took additional steps to further restrict disinformation. Following Facebook's debut of the "disputed" label

142. Adam Mosseri, "Addressing Hoaxes and Fake News," *Facebook Newsroom*, 15 December 2016, https://newsroom.fb.com/news/2016/12/news-feed-fyi-addressing-hoaxes-and-fake-news/.

143. "How is Facebook addressing false news through third-party fact-checkers?" Facebook Help Center, https://www.facebook.com/help/1952307158131536.

a year prior, the company found that many of the users who read the disputed information and warning label actually became *more* inclined to share the news.[144] In response, Facebook changed its approach and began shrinking the link in the newsfeed and including a number of related articles that debunked the news in question.[145]

The platform also increased its advertising controls, announcing that all political and issue-specific ads and advertisers would be vetted. Advertisers now need to have their identity and location confirmed by Facebook before receiving authorization to run ads on their platform. Furthermore, all political and issue-based advertisements are clearly labeled with "political ad" in the top corner, along with information about who funded the ad.[146]

By November 2018, Facebook had also dramatically scaled up its fact-checking operation. It had expanded its network of third-party fact-checking providers to 23 countries.[147] The company also began to implement machine learning tools to help these fact-checkers. The tools can spot warning signs in articles and help their human counterparts prioritize their efforts.[148]

To provide more context about news sources, the platform rolled out a new "information" feature in the bottom corner of news articles being shared. Users can click on the information icon and see additional details about the publication and author of the piece.[149]

Finally, Facebook announced its intention to expand its fact-checking service to photo and video content in order to warn users about doctored media that depict inaccurate events or media that is falsely attributed to people or events.[150]

In October 2018, Facebook launched its "Hunt for False News" blog to detail case studies and offer some transparency about how its platform is tackling disinformation.[151] During the 2018 US midterm elections, Facebook provided direct access to US state government officials to report posts containing false information about voting.[152]

---

144. Jeff Smith, Grace Jackson, and Seetha Raj, "Designing Against Misinformation," Medium Blog, 20 December 2017, https://medium.com/facebook-design/designing-against-misinformation-e5846b3aa1e2.
145. Josh Constine, "Facebook shrinks fake news after warnings backfire," ***TechCrunch***, May 2018, https://techcrunch.com/2018/04/27/facebook-false-news/.
146. Alex Hern, "New Facebook controls aim to regulate political ads and fight fake news," *The Guardian*, 06 April 2018, https://www.theguardian.com/technology/2018/apr/06/facebook-launches-controls-regulate-ads-publishers.
147. Jonathan Vanian, "Facebook Expanding Fact-Checking Project to Combat Fake News," *Fortune*, 21 June 2018, http://fortune.com/2018/06/21/facebook-fake-news-fact-checking/; and "Third-Party Fact-Checking on Facebook," Facebook Business, https://www.facebook.com/help/publisher/182222309230722?helpref=faq_content.
148. Constine, "Facebook shrinks fake news after warnings backfire."
149. Kif Leswing, "Facebook is finally launching a new feature to combat fake news, after six months of testing – and here's how it works," *Business Insider*, 03 April 2016, https://www.businessinsider.com/facebook-fake-news-about-this-article-feature-2018-4.
150. Lulu Chang and Eric Brackett, "Facebook outlines plans for combating false news," *Digital Trends*, 21 June 2018, https://www.digitaltrends.com/social-media/facebook-fight-fake-news/.
151. Julia Alexander, "Facebook introduces 'Hunt for False News' series in attempt to be transparent about misinformation," *The Verge*, 19 October 2018, https://www.theverge.com/2018/10/19/18001288/facebook-fake-news-hunt-spread-blog-election-misinformation.
152. Kevin Collier, "Facebook And Twitter Are Taking Steps To Help States Keep Bogus Election Information Off Their Sites," *Buzzfeed News*, 23 October 2018, https://www.buzzfeednews.com/article/kevincollier/facebook-twitter-false-information-election.

However, the impact of these actions is still unclear. Facebook's actions may now make it harder for Russian trolls to covertly influence politics abroad, but the company's new policy vetting advertisers still does not prevent foreign government-backed media outlets from *overtly* running influence campaigns, as evidenced by the ongoing proliferation of Chinese influence advertisements on Facebook.

Likewise, Facebook's unsuccessful experiment with flagging disputed news sources suggests that many people are undeterred from sharing articles even when the facts have been questioned by a third party. Removing contested articles could mitigate this risk, but permanently removing such contested news, rather than minimizing its appearance and newsfeed position, would open Facebook to criticism that it is censoring free speech and unfairly targeting political views. In March 2018, Mark Zuckerberg conveyed his strong reluctance to Facebook making those kinds of editorial decisions.[153]

On Facebook-owned Whatsapp, disinformation has been widespread. In India, for instance, a doctored video appearing to show a child kidnapping **went viral on WhatsApp, leading to more than 30 deaths from dozens of mob incidents** between 2017 and 2018.

Facebook's current approach towards disinformation leaves some vulnerabilities. One is on the platform's messaging apps. It is especially difficult for users to discern the validity and source of text content shared via private message. Further challenging matters, private messages are shared between users without the same level of fact-checking safeguards that takes place on a public newsfeed.

On Facebook-owned WhatsApp, which features an encrypted messaging service and has approximately 1.5 billion monthly users across the globe, disinformation and misinformation has been widespread.[154] In India, for instance, a doctored video appearing to show a child kidnapping went viral on WhatsApp, leading to more than 30 deaths from dozens of mob incidents between 2017 and 2018.[155] During Brazil's 2018 presidential election campaign, millions of dollars were spent by groups spreading disinformation on WhatsApp about the leftist candidate Fernando Haddad, who ultimately lost to his opponent Jair Bolsonaro.[156] A 2018 survey by Nieman Lab indicates that more than a third of Kenyans and South Africans, and almost 30 percent of Nigerians, have shared online fake news on WhatsApp, which is the dominant platform for mobile messaging in those countries.[157]

Facebook has recently taken steps to counter disinformation on messaging apps, but the opportunity for abuse remains significant. Mark Zuckerberg claimed in a 2018 *Vox* interview that the company's algorithms scan private message content to prevent sharing of harmful content, but the well-documented and ongoing proliferation of inflammatory anti-Rohingya content on

153. Kurt Wagner, "Mark Zuckerberg says he's 'fundamentally uncomfortable' making content decisions for Facebook," *Recode*, 22 March 2018, https://www.recode.net/2018/3/22/17150772/mark-zuckerberg-facebook-content-policy-guidelines-hate-free-speech.

154. Josh Constine, "WhatsApp hits 1.5 billion monthly users. $19B? Not so bad," *TechCrunch*, 31 January 2018, https://techcrunch.com/2018/01/31/whatsapp-hits-1-5-billion-monthly-users-19b-not-so-bad/.

155. Timothy Mclaughlin, "Disinformation is Spreading on WhatsApp in India - And It's Getting Dangerous," *The Atlantic*, 05 September 2018, https://www.theatlantic.com/international/archive/2018/09/fighting-whatsapp-disinformation-india-kerala-floods/569332/.

156. Tai Nalon, "Did WhatsApp Help Bolsonaro Win the Brazilian Presidency?" *Washington Post*, 01 November 2018, https://www.washingtonpost.com/news/theworldpost/wp/2018/11/01/whatsapp-2/?utm_term=.7a9e914f926d.

157. Herman Wasserman and Dani Madrid-Morales "New Data Suggests African Audiences See Significantly More Misinformation than Americans do," Nieman Lab, November 2018, http://www.niemanlab.org/2018/11/new-data-suggests-african-audiences-see-significantly-more-misinformation-than-americans-do/; and Daniel Funke, "Nigeria is the Next Battleground for Election Misinformation," The Poynter Institute, 30 November 2018, https://www.poynter.org/fact-checking/2018/nigeria-is-the-next-battleground-for-election-misinformation/.

Facebook Messenger in Myanmar suggests that its detection systems are not fail-safe.[158] In mid-2018, Facebook began testing new WhatsApp features, including labeling forwarded messages and limiting the number of forwards per user.[159] Labeling a message as "forwarded" still leaves the responsibility of identifying the authenticity of a message – and preventing its distribution if deemed false – squarely on the user.

A second vulnerability in Facebook's operations is its heavy reliance on outsourced, third-party fact-checking services staffed by human operators. Many of these groups are under-resourced and overwhelmed by the sheer volume of false or unverified content on the platform. For example, Facebook's fact-checking provider in the Philippines, Rappler, has been inundated with false news and struggling to keep up with the volume of disinformation created daily.[160]

Even when fact-checkers can handle the volume of content, fake news can easily go viral in the time between its creation and when fact-checkers are able to manually dispute the content and adjust its newsfeed ranking. If disinformation reaches a broad audience before its removal, it will have accomplished its intended purpose.

Nonetheless, Facebook's recent steps offer promise that the platform will be able to restrict the spread of disinformation. Some early research suggests that their efforts may be succeeding. Academics at New York University and Stanford University have analyzed the proliferation of more than 500 fake news websites and more than 10,000 fake news stories on platforms including Facebook between January 2015 and July 2018. They found that the spread of fake news on Facebook rose from 70 million engagements per month in early 2015 to 200 million engagements around the 2016 US presidential election. Then, the data peaked. The spread of fake news fell throughout 2017 and returned to approximately 70 million monthly engagements by early 2018.[161]

The reduction in Facebook's fake news engagement in 2017 and 2018 coincides with the company's efforts to counter disinformation, indicating that their tactics may be effective. Yet other factors might also explain this decline. For instance, neither 2017 nor 2018 had US elections of the same prominence as the 2016 presidential election, which may have lessened foreign incentives to meddle in the first place. Malicious actors may have also changed their tactics and sources to better evade detection in the latter stages of the study. These kinds of confounding factors, along with the proprietary nature of Facebook's data, complicates the ability to perform a comprehensive assessment and draw definitive conclusions.

158. Ezra Klein, "Mark Zuckerberg on Facebook's Hardest Year, and What Comes Next," *Vox*, 02 April 2018, https://www.vox.com/2018/4/2/17185052/mark-zuckerberg-facebook-interview-fake-news-bots-cambridge; and Steve Stecklow ,"Why Facebook is Losing the War on Hate Speech in Myanmar," *Reuters*, 15 August 2018, https://www.reuters.com/investigates/special-report/myanmar-facebook-hate/.

159. "Labeling Forwarded Messages," WhatsApp Blog, 10 July 2018, https://blog.whatsapp.com/10000645/Labeling-Forwarded-Messages; and "More Changes to Forwarding," WhatsApp Blog, 19 July 2018, https://blog.whatsapp.com/10000647/More-changes-to-forwarding.

160. Alexandra Stevenson, "Soldiers in Facebook's War On Fake News Are Feeling Overrun," *New York Times*, 09 October 2018, https://www.nytimes.com/2018/10/09/business/facebook-philippines-rappler-fake-news.html.

161. Hunt Allcott, Matthew Gentzkow, and Chuan Yu, "Trends in the Diffusion of Misinformation on Social Media," Stanford Institute for Economic Policy Research, October 2018, http://web.stanford.edu/~gentzkow/research/fake-news-trends.pdf.

Countering disinformation at Twitter

Like Facebook, Twitter has grappled with the spread of disinformation through bots, fake accounts, and advertisements. In some respects, Twitter is uniquely vulnerable to coordinated and viral disinformation campaigns. While most Facebook users' newsfeed activity is tied to their personal networks, Twitter is a platform where news and commentary is often shared widely through issue-specific hashtags, public comment threads, and influencer accounts with large followings.

Researchers from the Massachusetts Institute of Technology analyzed thousands of stories on Twitter that were tweeted millions of times between 2016 and 2017 and found that disinformation often reached a broader audience than the truth. For example, the top one percent of fake news stories typically reached between 1,000 and 100,000 people, whereas true stories rarely exceeded an audience of one thousand. In line with similar studies, the researchers concluded that people were more likely to share fake news because these stories generated more reactions of surprise and disgust.[162]

On Twitter, **the top one percent of fake news stories typically reached between 1,000 and 100,000 people**, whereas true stories rarely exceeded an audience of one thousand.

Twitter bots play an important role in disseminating disinformation. Researchers analyzing the spread of 400,000 articles via 14 million messages on Twitter between 2016 and 2017 found that bots were an important vehicle for sharing content that had "low-credibility sources."

Bots regularly share low-credibility content in the first few seconds of its publication, thereby increasing the number of impressions and consequently its chances of going viral. Six percent of the identified bots accounted for 31 percent of the spread of low-credibility content, typically within the first 10 seconds after an article is published.[163] Furthermore, a Knight Foundation study of 700,000 Twitter accounts linked to disinformation during the 2016 US election also found that the majority of these accounts were either fully or semi-automated bots.[164]

Recent estimates suggest that the number of bots on the platform increased significantly between 2014 and 2017. Twitter estimated in 2014 that between five and 8.5 percent of its user base consisted of bots. Several years later, in March 2017, research by academics from Indiana University and the University of Southern California indicated that the prevalence of bots was significantly higher: 15 percent or more of the platform's total accounts.[165]

Twitter began taking stronger action to counter this threat in October 2017, when the company announced a new set of policies to increase advertising transparency. The initiative, which labels political ads with the organization that purchased the content, also allows users to report inappropriate

162. Soroush Vosoughi, Deb Roy, Sinan Aral, "The spread of true and false news online," *Science*, 09 March 2018, http://science.sciencemag.org/content/359/6380/1146.

163. Chengcheng Shao et al, "The spread of low-credibility content by social bots," *Nature Communications*, 20 November 2018, https://www.nature.com/articles/s41467-018-06930-7.

164. Knight Foundation, "Disinformation, 'Fake News' and Influence Campaigns on Twitter," 04 October 2018, https://knightfoundation.org/reports/disinformation-fake-news-and-influence-campaigns-on-twitter.

165. Onur Varol et al, "Online Human-Bot Interactions: Detection, Estimation, and Characterization," Indiana University and University of Southern California, 27 March 2017, https://arxiv.org/pdf/1703.03107.pdf.

ads and to inform Twitter when they see ads they do not like. Twitter also launched a "Transparency Center," which allows users to search political ad campaigns and view additional details including ad creative, ad campaign history, targeted demographics and dollar amounts spent by the advertiser.[166]

In May 2018 Twitter announced that it was using machine learning and artificial intelligence to identify traits associated with accounts engaged in trolling or coordinated disinformation attacks. Individuals, for instance, who sign up for multiple Twitter accounts simultaneously would be flagged.

While content from these accounts is not removed, it is deprioritized in search and conversation results. Twitter found in early testing that reports of abuse in conversations dropped eight percent, and similar reports from search content dropped four percent.[167]

A month after Twitter announced its use of these AI tools, it also completed the acquisition of Smyte, a technology company focused on countering spam, fraud, and abuse online. Twitter incorporated Smyte's existing tools into its platform, including Smyte's labeling system that automatically identifies potentially malicious activity. It does this based on a number of signals and account relationships, and then flags the activity for internal review.[168]

In May and June of 2018, Twitter also began accelerating the pace with which it suspended suspicious accounts. In those two months alone, Twitter suspended 70 million accounts, twice the rate of suspensions compared to fall 2017.[169]

This was far from a trivial move for Twitter. Removing such a large volume of accounts impacted its active user metrics, which play an important role in the financial valuation of social media companies. Twitter's July 2018 quarterly earnings announced negative user growth as a result of their account purge, and the company's stock promptly fell 21%.[170] (This clear tradeoff between suspending fake accounts and the company's valuation may help to explain Twitter's initial reluctance to more proactively remove suspicious accounts.)

Nevertheless, Twitter's overall record on countering disinformation appears mixed at best. In mid-2016, Twitter reportedly offered the Kremlin-backed Russia Television Network a 15 percent share of its US election advertising in exchange for three million dollars.[171]

166. @brucefalck, "New Transparency for Ads on Twitter," Twitter Blog, 24 October 2017, https://blog.twitter.com/official/en_us/topics/product/2017/New-Transparency-For-Ads-on-Twitter.html.
167. @delbius and @gasca, "Serving Healthy Conversation," Twitter Blog, 15 May 2018, https://blog.twitter.com/official/en_us/topics/product/2018/Serving_Healthy_Conversation.html.
168. @twittersafety, "Continuing our commitment to health," Twitter Blog, 21 June 2018, https://blog.twitter.com/official/en_us/topics/company/2018/CommitmentToHealth.html.
169. Craig Timberg and Elizabeth Dwoskin, "Twitter is sweeping out fake accounts like never before, putting user growth at risk," *Washington Post*, 06 July 2018, https://www.washingtonpost.com/technology/2018/07/06/twitter-is-sweeping-out-fake-accounts-like-never-before-putting-user-growth-risk/?utm_term=.ccd2163a4107.
170. "Twitter stock plunges 21% after earnings show effects of fake-account purge," *Marketwatch*, 28 July 2018, https://www.marketwatch.com/story/twitter-shares-slide-16-after-fake-account-purge-new-rules-in-europe-2018-07-27.
171. Alex Kantrowitz, "Twitter Offered Russian Television Network RT 15% Of Its Total Share Of US Elections Advertising," *Buzzfeed News*, 01 November 2017, https://www.buzzfeednews.com/article/alexkantrowitz/twitter-offered-rt-15-of-its-total-share-of-us-elections.

Furthermore, the Knight Foundation study found that more than 80 percent of the 700,000 Twitter accounts linked to disinformation in the 2016 election were still active in spring 2018, and many were continuing to produce millions of tweets. They also found that nearly 90 percent of Twitter disinformation was traced back to the same 50 fake news sites, many of which remained leading sources of disinformation on Twitter for many months after the 2016 election. Conspiracy news sites received approximately 13 percent the amount of Twitter links as a comparative set of national news sites, indicating that the spread and consumption of fake news was extensive.[172]

With more than **1.8 billion monthly users, YouTube is a powerful conveyor of video news**, without many of the **traditional gatekeepers that curate content** on television news channels.

The NYU and Stanford researchers confirmed the trend. While they found that fake news was decreasing on Facebook after 2016, their analysis concluded it was rising on Twitter: Fake news accounted for two million shares per month in 2015, about 4.5 million shares per month at the end of 2016, and about six million shares per month in mid-2018.[173]

In August 2018, Twitter CEO Jack Dorsey expressed hesitation about taking forceful action against perceived fake news. In a public interview, he raised concerns about Twitter becoming the arbiter of truth by projecting the company's biases onto the conversations that are allowed on, and banned from, its platform.

Dorsey further highlighted the logistical challenge of reviewing all content on Twitter for signs of fake news – noting the large volume of personnel resources needed for such an undertaking – to justify reviewing content only once it has been flagged by users.[174] Dorsey's statements, together with his company's mixed record on tackling known sources of disinformation, raise some questions about how proactive Twitter will be in countering this threat going forward.

Countering disinformation at Google

Malicious actors have spread disinformation on Google's platform in a number of ways. While Google-owned YouTube has hosted fake video content, there have also been disinformation ads on Google's ad platform, and bad actors have manipulated Google's search results to prominently feature fake news sites.

YouTube is especially prone to disinformation, and the platform has faced strong criticism for its algorithms that determine trending content and viewers' personalized recommendations. With more than 1.8 billion monthly users, YouTube is a powerful conveyor of video news, without many of the traditional gatekeepers that curate content on television news channels.[175] One consequence is that YouTube's trending list is vulnerable to sensationalized content, where algorithms optimize for virality over quality or accuracy. The platform's algorithms can also inadvertently promote disinformation through personalized "up next" video recommendations.

172. Knight Foundation, "Disinformation, 'Fake News' and Influence Campaigns on Twitter."

173. Hunt Allcott, Matthew Gentzkow, and Chuan Yu, "Trends in the Diffusion of Misinformation on Social Media," Stanford Institute for Economic Policy Research, October 2018, http://web.stanford.edu/~gentzkow/research/fake-news-trends.pdf.

174. Jon Fingas, "Jack Dorsey Explains Why Twitter Is Reluctant to Fight Fake News," *Engadget*, 19 August 2018, https://www.engadget.com/2018/08/19/jack-dorsey-explains-twitter-reluctance-to-fight-fake-news/.

175. Ben Gilbert, "YouTube now has over 1.8 billion users every month, in spitting distance of Facebook's 2 billion," *Business Insider*, 04 May 2018, https://www.businessinsider.com/youtube-user-statistics-2018-5.

These recommendations are critical to YouTube's success; the company claims that 70 percent of its more than one billion hours of daily video consumption is driven by its playlist recommendations.[176] Algorithms automatically generate suggested content intended to maximize user engagement, which means serving users content that mimics or appeals to their past behavior. If a viewer consumes conspiratorial or other misleading content on YouTube, its algorithms will recommend similar content, which can reinforce the viewers' belief that the conspiracy is, in fact, real.

What is more concerning, however, is that viewing questionable content has not been a prerequisite for receiving questionable recommendations. One does not have to be a conspiracy theorist to be fed conspiracy theories. A 2018 *Wall Street Journal* investigation found that viewing mainstream political content on YouTube, across the political spectrum, led to content suggestions with more extreme and fringe viewpoints than displayed in the original video.[177]

Perhaps with these concerns in mind, Google has amplified its effort over the past two years to counter disinformation on its platforms. Following the 2016 election, Google changed its advertising policies to ban known disinformation websites from using Google's "AdSense" program which enables companies to monetize websites through ad placements.[178] In January 2017, Google banned a group of 200 publishers from its AdSense network who violated these policies.[179]

Building upon these efforts, Google announced a partnership with the International Fact Checking Network in October 2017 to expand a global network of fact-checkers and provide free tools to support its fact-checking efforts.[180] That same month, however, Google was found to be unknowingly serving fake news ads on the fact-checking websites Politifact and Snopes, which served as an awkward reminder of the ongoing challenges faced by Google and host websites in dealing with disinformation in advertisements.[181]

Google took steps in 2017 to improve the transparency and quality of its search engine results. It introduced a "fact-check" feature that would display third-party publications' fact-checking assessments underneath Google's results for frequently-searched public claims.[182] However, in January 2018, Google announced it was temporarily suspending its fact-checking feature after conservative media outlets raised concerns that the feature was being selectively – and in some cases, incorrectly – applied to conservative media reporting.[183]

176. Jack Nicas, "How YouTube Drives People to the Internet's Darkest Corners," *Wall Street Journal*, 07 February 2018, https://www.wsj.com/articles/how-youtube-drives-viewers-to-the-internetInternets-darkest-corners-1518020478.
177. Nicas, "How YouTube Drives People to the Internet's Darkest Corners."
178. Jack Nicas, "Google to Bar Fake-News Websites From Using Its Ad-Selling Software," *Wall Street Journal*, 04 November 2016, https://www.wsj.com/articles/google-to-bar-fake-news-websites-from-using-its-ad-selling-software-1479164646.
179. Tess Townsend, "Google has banned 200 publishers since it passed a new policy against fake news," *Recode*, 25 January 2017, https://www.recode.net/2017/1/25/14375750/google-adsense-advertisers-publishers-fake-news.
180. Erica Anderson, "Building trust online by partnering with the International Fact Checking Network," Google Blog, 26 October 2017, https://www.blog.google/outreach-initiatives/google-news-initiative/building-trust-online-partnering-international-fact-checking-network/.
181. Daisuke Wakabayashi and Linda Qiu, "Google Serves Fake News Ads in an Unlikely Place: Fact-Checking Sites,", *New York Times*, 17 October 2017, https://www.nytimes.com/2017/10/17/technology/google-fake-ads-fact-check.html.
182. Justin Kosslyn and Cong Yu, "Fact Check now available in Google Search and News around the world," Google Blog, 07 April 2017, https://blog.google/products/search/fact-check-now-available-google-search-and-news-around-world/.
183. Daniel Funke, "Google suspends fact-checking feature over quality concerns," Poynter Institute, 19 January 2018, https://www.poynter.org/news/google-suspends-fact-checking-feature-over-quality-concerns.

Social media's business model, **by promoting clicks, views, and shares, can be easily weaponized by bad actors to spread disinformation** – and many might argue that it is optimized to do so.

For searches about media publications, Google added a "Knowledge Panel" section dedicated to providing additional information about the publication, including its most commonly-covered topics, as well as previous awards won for its reporting.[184] Google also announced that news sites that hid or misrepresented their country of origin would no longer appear in its search results.[185]

As part of a $300 million "Google News Initiative" announced in March 2018, Google committed $10 million to an anti-disinformation media literacy effort, which would produce educational content in partnership with YouTube stars to help educate the public on how to identify fake news online.[186]

Google also announced in July 2018 that YouTube would more heavily promote "authoritative" news sources in the wake of breaking news. It also began sharing short text previews summarizing these breaking news stories in the search results, particularly during the time gap between a story first breaking and when traditional media sources are able to release verified, high-quality video content.[187] This strategy is intended to curb the spread of disinformation and misinformation clips that can often go viral in the immediate aftermath of a breaking news story.

Furthermore, in response to increasing public scrutiny over its recommendation algorithms, YouTube announced in early 2019 that it would begin adjusting its algorithms to reduce its recommendations of harmful or purposefully misleading content.[188]

The structural barriers to tackling disinformation on tech platforms

The past several years have shown that there are times when social media companies will not – or cannot – adequately police their own platforms against all forms of disinformation.

Eliminating every manifestation of disinformation is an extraordinarily difficult challenge, considering the sheer number of bad actors, huge volume of daily content from hundreds of millions of users, and quickly-changing technology landscape. In many cases, these platforms' most meaningful efforts against disinformation have occurred in response to public outcries. Yet, their reactionary efforts have not succeeded in preventing the subsequent, more evolved disinformation strategies by foreign states. There are always new gaps to be exploited.

184. Ranna Zhou, "Learn more about publishers on Google," Google Blog, 07 November 2017, https://blog.google/products/search/learn-more-about-publishers-google/.

185. Jon Fingas, "Google won't show news from sites that hide their country of origin," *Engadget*, 16 December 2017, https://www.engadget.com/2017/12/16/google-bans-news-sites-which-hide-country-of-origin/.

186. Shan Wang, "Google announces a $300M 'Google News Initiative'," NiemanLab, 20 March 2018, http://www.niemanlab.org/2018/03/google-announces-a-300m-google-news-initiative-though-this-isnt-about-giving-out-grants-directly-to-newsrooms-like-it-does-in-europe/.

187. Barbara Ortutay ,"YouTube is cracking down on 'fake news' with new text previews," *Associated Press*, 09 July 2018, https://www.usatoday.com/story/tech/2018/07/09/youtube-cracks-down-fake-news/769861002/.

188. "Continuing our work to improve recommendations on YouTube," Official blog, YouTube, January 25, 2019, https://youtube.googleblog.com/2019/01/continuing-our-work-to-improve.html.

Social media's business model, by promoting clicks, views, and shares, can be easily weaponized by bad actors to spread disinformation – and many might argue that it is optimized to do so.[189] Research from New America reiterates that the central problem of disinformation is not adversarial state meddling or even just one platform's vulnerabilities, but rather that the entire social media industry has been built to leverage sophisticated technology to aggregate user attention and sell advertising.[190]

As such, strategies to counter disinformation must consider a number of digital marketing features, including: behavioral data tracking that enables precision targeting; online ad buying to reach and impact certain audiences; search engine optimization that tricks algorithms and dominates search results; social media management services that preconfigure messages for select audiences across multiple media channels; and marketing AI that improves behavioral data tracking, audience segmentation, message targeting, and campaign management.[191]

Furthermore, although it is tempting to simply implicate the platforms' business models as the primary conveyor of online disinformation, there are also the social and psychological factors that independently and concurrently enable its spread. Consider that WhatsApp, Facebook's messaging platform, had yet to implement an effective advertising model for its 1.5 billion users through the end of 2018, but has nonetheless grappled with some of the most pervasive fake news campaigns globally.[192] On messaging apps largely devoid of advertising, bad actors have been able to spread disinformation by leveraging influencers and trusted networks to widely share inflammatory content over message groups.

Efforts to counter disinformation can also have unintended consequences, such as Facebook's experiment adding warning labels to disputed content - which actually increased the spread of disinformation. Other efforts have invited critiques that the platforms are stifling free speech or are politically biased.

Recognizing that the challenge posed by disinformation is massive and transcends any single technical vulnerability, other organizations like academia, non-governmental organizations, and technology startups have stepped in to develop solutions to counter certain aspects of online disinformation.[193] These organizations, though, face the problem of information asymmetry. They simply do not have the same level of access to relevant platform-specific algorithms and raw user data as the

189. See Anthony Nadler, Matthew Crain, and Joan Donovan, "Weaponizing the Digital Influence Machine," Data&Society, 17 October 2018, https://datasociety.net/output/weaponizing-the-digital-influence-machine/; Douglas Guilbeault, "Digital Marketing in the Disinformation Age," ***Journal of International Affairs***, Columbia University School of International and Public Affairs, 17 September 2018, https://jia.sipa.columbia.edu/digital-marketing-disinformation-age; and Dipayan Ghosh and Ben Scott, "Russia's Election Interference is Digital Marketing 101," *The Atlantic*, 19 February 2018, https://www.theatlantic.com/international/archive/2018/02/russia-trump-election-facebook-twitter-advertising/553676/.

190. Dipayan Ghosh and Ben Scott, "#DigitalDeceit - The Technologies Behind Precision Propaganda on the Internet," 23 January 2018, New America, https://www.newamerica.org/public-interest-technology/policy-papers/digitaldeceit/.

191. Ghosh and Scott, "#DigitalDeceit."

192. Parmy Olson, "Facebook's Revenue Dreams for WhatsApp Are Under Threat – From Old-Fashioned SMS," *Forbes*, 21 November 2018, https://www.forbes.com/sites/parmyolson/2018/11/21/facebooks-revenue-dreams-for-whatsapp-are-under-threatfrom-old-fashioned-sms/#372cb358556e.

193. See, for example, Center for American Progress et al, "Recommended Internet Company Corporate Policies and Terms of Service to Reduce Hateful Activities," 2018, https://uploads-ssl.webflow.com/5bba6f4828dfc3686095bf6b/5bd0e36186e28d35874f0909_Recommended%20InternetInternet%20Company%20Corporate%20Policies%20%20Terms%20of%20Service_final-10-24.pdf.

platforms themselves. For third-party organizations to develop truly effective ideas and solutions to pieces of the online disinformation puzzle, they need to collaborate with platforms and work jointly towards solutions.

Taking responsibility

Who then bears responsibility for countering disinformation?

In some ways, the answer appears simple. When states like Russia or Iran spread disinformation on Facebook or Twitter, they are not doing so to attack Facebook or Twitter. They are doing it to undermine geopolitical adversaries, including the United States. Governments, then, seem to bear the ultimate responsibility for defending their nations against this kind of disinformation.

However, that answer obscures a major complication: the battleground rests firmly in private hands.

In the absence of clear delineations of responsibility, **a reasonable next step could involve greater collaboration between technology companies and governments.**

This would suggest that there is a greater role for governments to play in engaging with and regulating social media companies. After all, online platforms, while well-resourced both financially and technically to wage this battle, do not necessarily have perfectly-aligned incentives with governments who are seeking to guard against foreign meddling. Nor are they necessarily capable of defending against every effort from sophisticated hostile actors.

On the other hand, significant government involvement carries its own risks, including the potential for impinging upon freedom of expression and outright censorship.[194] However, certain tailored regulations may avoid such limitations. For example, Guillaume Chaslot, a former YouTube software engineer, has suggested holding technology companies legally liable for their algorithmic recommendations, as opposed to every piece of content they host.[195] Such an approach could protect freedom of expression while still holding social media companies accountable, and incentivized, to prevent their platforms from recommending disinformation-related content.

In the absence of clear delineations of responsibility, a reasonable next step could involve greater collaboration between technology companies and governments. A productive public-private relationship would enable transparent information sharing, fact-finding, and the development and deployment of targeted solutions meant to quickly counter foreign disinformation online.

A number of successful public-private structures already exist that can serve as a model for countering disinformation. One such model to consider emulating is the Global Internet Forum to Counter Terrorism (GIFCT), led by Facebook, Google, Microsoft, and Twitter. Formed in 2017, the GIFCT is an initiative working with governments, multilateral organizations including the UN, NGOs, and academia to curb the online spread of terrorist content in a manner that protects human rights, freedom of expression,

194. Darrell M. West, "How to combat fake news and disinformation," Brookings Institution, 18 December 2017, https://www.brookings.edu/research/how-to-combat-fake-news-and-disinformation/.

195. Kevin Roose, "YouTube unleashed a conspiracy theory boom. Can it be contained?" *New York Times*, 19 February 2019, https://www.nytimes.com/2019/02/19/technology/youtube-conspiracy-stars.html.

user privacy, and the role of journalism.[196] As one example of their collaborative work, by mid-2018 GIFCT member companies had added nearly 100,000 links of terrorism-related content to a shared database that allows participating companies to collectively block the material before it is posted.[197]

Another potential model comes from the world of financial crimes enforcement, where several frameworks promote cooperation between governments and the financial sector to better identify and disrupt money laundering and terrorist financing. One prominent example is the inter-governmental Financial Action Task Force (FATF), which encourages information sharing between financial institutions, law enforcement authorities, and governments. The FATF works to identify country-level vulnerabilities, promote regulatory reform, and leverage new technologies to confront money laundering and terrorist financing across its 37 member states.[198]

Given the numerous actors that shape and are shaped by the information and digital landscape, addressing disinformation will require ongoing and open cooperation.[199] There is no single solution or silver bullet to address this complex problem. However, social media and technology companies are well-placed to lead these efforts, in collaboration with governments and other partners.

196. "Global Internet Forum to Counter Terrorism," https://www.gifct.org/about/.

197. Dave Lee, "Tech firms hail 'progress' on blocking terror," BBC News, 8 June 2018, https://www.bbc.com/news/technology-44408463.

198. FATF Guidance – Private Sector Information Sharing, http://www.fatf-gafi.org/fr/publications/recommandationsgafi/documents/guidance-information-sharing.html?hf=10&b=0&s=desc(fatf_releasedate)

199. "A multi-dimensional approach to disinformation: Report of the independent High level Group on fake news and online disinformation," European Commission, 12 March 2018, https://ec.europa.eu/digital-single-market/en/news/final-report-high-level-expert-group-fake-news-and-online-disinformation.

6. Knowledge gaps and future technology challenges

The challenge of confronting disinformation will continue to change as old technologies evolve and new ones develop. These changes may widen existing gaps between the threat of disinformation and the ability to counter it. To understand these challenges, this section outlines three broad categories that impact efforts against disinformation.

First are technology gaps. Many observers classify the modern disinformation environment as an arms race in which researchers, technologists, and governments scramble to develop tools to detect, counter, and keep pace with nefarious actors' methods and activities. This environment is characterized by a wide availability of sophisticated technologies that, until recently, were concentrated in leading tech companies or research labs.

The second set of challenges is structural. These relate to the economic incentives of developing counter-disinformation technology, the dearth of available data sets to train machine-learning tools, and the slow rate of adoption of existing tools.

The third and final category of challenges relate to the gap in understanding exactly how technologies – such as AI – are evolving, and with it, the threat from disinformation.

What follows is a snapshot of areas where technology, structural, and knowledge gaps currently hinder the detection and prevention of online disinformation. It is by no means exhaustive, but aims to highlight several important categories that warrant attention over the coming years.

Gaps in technology

Bots

Bots are becoming more sophisticated, thanks in part to technology developed to exploit the lucrative online marketing and advertising markets.[200] Detecting spambots on Twitter based on syntax, semantics, or network features is effective. However, detecting next generation political bots that do

200. Samuel Woolley and Marina Gorbis, "Social Media Bots Threaten Democracy. But We Are Not Helpless," *The Guardian*, 16 October 2017, https://www.theguardian.com/commentisfree/2017/oct/16/bots-social-media-threaten-democracy-technology?CMP=twt_gu.

not just repeat or retweet information but actually become intelligent and capable of producing content on their own is a challenge.

Most of the work on political bots has focused on short periods of time, in a specific political context, and on Twitter. Questions remain about the average lifespan of a bot, the transferability from one country or political context to another, and the efficacy of building algorithms to detect bots on platforms other than Twitter.[201]

Photos and videos

Detecting altered photos and videos at scale is difficult, and rapidly advancing AI and deep learning technology is making synthetic media (manipulated or artificially-created video and audio content) easier to produce.

Images are easy to manipulate but more difficult to detect with the currently available image analysis and forensic tools. The Defense Advanced Research Projects Agency's "Media Forensics" program is undertaking an effort to develop and deploy such tools on a platform that will automate the assessment of an image's integrity. Commercial solutions are also coming to market. One company called Truepic is releasing an image and video forensics tool in mid-2019 that authenticates media by scanning for any abnormalities that would indicate manipulation.[202]

Further work will be needed to stay abreast of advancements in doctoring videos.[203] As AI technology progresses, synthetic video and audio will appear increasingly authentic to the public and will become significantly easier to manufacture. This will lead to the migration of disinformation content from being largely "static" (memes, fake articles) to "dynamic" (video and audio).[204] For example, the video mapping of one person's face onto another, termed a "deep fake," is already widely available through public apps. Video to video synthesis technology can create realistic looking artificial video content based on a set of inputs.[205]

While reverse image search tools exist, robust reverse video search tools are similarly needed to detect synthetic video content. Reverse image search can be used to identify the source of an image online by pointing to where else it has appeared, which helps people verify the origin of an image quickly. Google Reverse Image Search is one such tool. However, reverse video search tools are limited in their functionality, relying on thumbnails or dissected portions of videos.

201. Joshua A. Tucker et al, "Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature," William + Flora Hewlett Foundation, 19 March 2018, https://hewlett.org/library/social-media-political-polarization-political-disinformation-review-scientific-literature/.

202. J.J. McCorvey, "This image-authentication startup is combating faux social media accounts, doctored photos, deep fakes, and more," *Fast Company*, 19 February 2019, https://www.fastcompany.com/90299000/truepic-most-innovative-companies-2019.

203. Elizabeth Gibney, "The scientist who spots fake videos," *Nature*, 06 October 2017, https://www.nature.com/news/the-scientist-who-spots-fake-videos-1.22784.

204. Alina Polyakova, "Weapons of the weak: Russia and AI-driven asymmetric warfare," Brookings' Artificial Intelligence and Emerging Technologies Initiative, 15 November 2018, https://www.brookings.edu/research/weapons-of-the-weak-russia-and-ai-driven-asymmetric-warfare/.

205. Polyakova, "Weapons of the weak: Russia and AI-driven asymmetric warfare."

A Gartner report predicts that by **2020, the abilities of AI to generate counterfeit media will surpass those of AI to identify such media**. Relying on algorithms alone to solve algorithm-driven disinformation may therefore be insufficient.

Going forward, as AI algorithms learn to better imitate reality through deep fake and synthetic videos, it will become increasingly difficult for algorithms to detect fakes. The gap between AI's ability to create rather than counter synthetic media is expected to widen. A Gartner report predicts that by 2020, the abilities of AI to generate counterfeit media will surpass those of AI to identify such media.[206] Relying on algorithms alone to solve algorithm-driven disinformation may therefore be insufficient.[207]

Fact-checking

Fact-checking initiatives have proliferated in recent years, but automated fact-checking is still a nascent area of development. The possibilities range from augmenting human fact-checkers with databases of source material to annotating content with fact-checks in news articles. Ongoing hurdles include how to teach computers to discern the parts of a sentence that should be fact-checked, and how to build databases with enough content to, for example, check a politician's pronouncements against their prior statements.

The bigger challenge, however, is how to reduce the resonance and reach of contested information. Facebook's failed experiment labeling false information as "disputed," which perversely caused more sharing of the flagged content, suggests that merely labeling disputed information as such does not lead to a reduction in its public appeal or organic spread.[208]

Structural challenges

Volume of content

In cases where solutions to disinformation are available, existing structural challenges can impede their effectiveness. The sheer magnitude of content and platforms is perhaps one of the biggest obstacles hindering monitoring and detection.

Even if platforms had the capability to detect synthetic images or altered videos and regulations forced them to use it, the enormous and steadily growing volume of content being uploaded raises questions about the ability of these platforms to effectively monitor all of it.[209] Hundreds of hours of video are uploaded to YouTube every minute.[210] Additionally, the rapid proliferation of platforms amplifies this problem. For example, a UK Home Office press release noted that in 2017, ISIS used more

206. Kasey Panetta ,"Gartner Top Strategic Predictions for 2018 and Beyond," Gartner, Inc., 03 October 2017, https://www.gartner.com/smarterwithgartner/gartner-top-strategic-predictions-for-2018-and-beyond/.

207. Chris Meserole and Alina Polyakova "Disinformation Wars," *Foreign Policy*, 25 May 2018, https://foreignpolicy.com/2018/05/25/disinformation-wars/.

208. Jeff Smith, Grace Jackson, and Seetha Raj, "Designing Against Misinformation," Medium Blog, 20 September 2017, https://medium.com/facebook-design/designing-against-misinformation-e5846b3aa1e2.

209. "Fighting Fake News: Workshop Report," The Information Society Project at Yale Law School and the Floyd Abrams Institute for Freedom of Expression, 07 March 2017, https://law.yale.edu/system/files/area/center/isp/documents/fighting_fake_news_-_workshop_report.pdf.

210. "Hours of video uploaded to YouTube every minute as of July 2015," The Statistics Portal, https://www.statista.com/statistics/259477/hours-of-video-uploaded-to-youtube-every-minute/.

than 400 unique platforms to distribute content. In the latter half of 2017 alone, ISIS used 145 new platforms for this purpose.[211]

Encryption

The growing adoption of encrypted and private messaging apps poses various challenges for countering disinformation. Rumors and disinformation that spread through a wide network of private group chats cannot be easily detected or fact-checked in real time, and certainly not by humans. Compounding matters, the private and encrypted nature of these apps prevents the platforms from publicly flagging content as false, widely disseminating corrections, or removing the objectionable content from message groups. Some experts suggest, however, that the growing use of encrypted messaging apps could also decrease the effectiveness of certain types of disinformation, as it will become harder to target select audiences whose identifying data is no longer freely available.[212]

Verification tools

A limited number of social media verification tools exist, such as reverse image search, but adoption of these tools remains a challenge. The International Center for Journalists conducted a study that found that 71 percent of journalists use social media to find stories but only 11 percent use any social media verification tools.[213] Platforms, too, will require convincing to adopt approaches like a uniform standard for imprinting or watermarking videos with their digital origin. As one report noted, "Even if an effective detection method emerges, it will struggle to have broad impact unless the major content distribution platforms, including traditional and social media, adopt it as a screening or filtering mechanism."[214]

Access to data

Large data sets are critical to train the machine learning tools used in counter-disinformation efforts, but researchers lack access to such data. Companies rarely share their data, and validated, empirical data on bots and trolls is difficult to find as their creators typically remain anonymous. This is one reason why machine learning alone is often insufficient and must be augmented with human review.[215]

There are available data sets that can be used to train generative adversarial networks on detecting synthetic images of faces, but more data sets are needed for other areas of forensic inquiry.[216] As one

211. Government of the United Kingdom Home Office and The Rt Hon Amber Rudd, Press Release, "New technology revealed to help fight terrorist content online," 13 February 2018, https://www.gov.uk/government/news/new-technology-revealed-to-help-fight-terrorist-content-online.

212. Interview with subject matter expert Sean Murphy, 12 January 2019.

213. Nic Dias, "The Big Question: How Will 'Deepfakes' and Emerging Technology Transform Disinformation?" National Endowment for Democracy, 01 October 2018, https://www.ned.org/the-big-question-how-will-deepfakes-and-emerging-technology-transform-disinformation/.

214. "Disinformation on Steroids," Council on Foreign Relations Digital and Cyberspace Policy Program, 16 October 2018, https://www.cfr.org/report/deep-fake-disinformation-steroids.

215. V.S. Subrahmanian et al, "The DARPA Twitter Bot Challenge," (On file with Cornell University), 20 January 2016, https://arxiv.org/pdf/1601.05140.pdf.

216. Andreas Rössler et al, "FaceForensics: A Large-scale Video Dataset for Forgery Detection in Human Faces," (On file with Cornell University), 24 March 2018, https://arxiv.org/abs/1803.09179.

organization noted, detection is currently easier than forgery so long as training data shows new types of faked images, audio, and video.[217] The pace of technological advancements also limits how quickly new tools can be developed to confront the challenges of disinformation. For example, new products may require specialized AI chips that can handle complex tasks but require new materials and production techniques.[218]

Furthermore, mandatory requirements for disinformation monitoring that require expensive or complex solutions can raise the barriers of entry to young, innovative companies and entrench the position of incumbents. It is worth noting, however, that some parties have developed solutions with the intent of sharing them with smaller companies – the UK government and ASI Data Science are doing so with technology they developed in collaboration to detect terrorist content online.[219]

Gaps in knowledge

To better understand disinformation's impact on people, additional research is needed to understand if, how, and when exposure to disinformation influences or changes the recipient's beliefs. To separate signal from noise, it is important to differentiate between 1) disinformation that is inauthentic and thus ineffective; 2) disinformation that simply appeals to preexisting prejudices without changing its target audience's views; and 3) disinformation that successfully alters viewpoints, discourse, and decisions. However, making these distinctions is not intended to dismiss the risks of disinformation "noise." A greater understanding of whether, and to what extent, seemingly inauthentic disinformation "noise" still undermines the public's trust in online media will help researchers better gauge and analyze the risk from all forms of disinformation, both sophisticated and unsophisticated.

On the technology side of the equation, experts continue to lack an understanding of why complex AI products behave as they do. Neural networks underpin many of the tools people use every day. Yet their inner workings are sometimes beyond comprehension.[220]

Without understanding the AI that adversaries use to create and disseminate disinformation, methods to counter such techniques may be difficult to generate. Research must also consider the potentially nefarious uses for ostensibly innocuous AI. For example, Microsoft's Xiaoice has been a popular chatbot in China since 2014 and is designed to fulfill human needs for communication, affection, and social belonging.[221] AI like Xiaoice could conceivably be trained by authoritarian governments to circumvent certain conversations or propagate certain ideas that fall within the realm of disinformation.

---

217. Sam Gregory, "Deepfakes and Synthethic Media: What Should We Fear? What Can We Do?" Witness Blog, July 2018, https://blog.witness.org/2018/07/deepfakes/.
218. Cade Metz, "Chips Off the Old Block: Computers Are Taking Design Cues From Human Brains," ***New York Times***, 16 September 2017, https://www.nytimes.com/2017/09/16/technology/chips-off-the-old-block-computers-are-taking-design-cues-from-human-brains.html.
219. Government of the United Kingdom Home Office and The Rt Hon Amber Rudd, Press Release, "New technology revealed to help fight terrorist content online," 13 February 2018, https://www.gov.uk/government/news/new-technology-revealed-to-help-fight-terrorist-content-online.
220. Will Knight, "The Dark Secret at the Heart of AI," *MIT Technology Review*, 11 April 2017, https://www.technologyreview.com/s/604087/the-dark-secret-at-the-heart-of-ai/.
221. Li Zhou et al, "The Design and Implementation of Xiaoice, an Empathetic Social Chatbot," (On file with Cornell University), 21 December 2018, https://arxiv.org/abs/1812.08989.

While not yet a reality, the anticipation of worst-case scenarios like these and others are what prompted researchers with OpenAI to scale back their release of a machine learning system that generated text based on brief writing prompts. The narratives produced by the fully trained model were alarming in their ability to mimic human writing, leading its developers to release only a small model for research purposes.[222]

Other efforts to develop virtual digital influencers that are completely computer-generated also pose interesting questions about how such influencers might be subverted to push malicious stories and ideas to their online followers.[223] One need only consider the example of Microsoft's "Tay" chatbot on Twitter that was active for only a day before it began mimicking the vitriolic racism and sexism it encountered on the platform.[224]

Living with disinformation

The **use of disinformation as a tool for influence and obfuscation will never cease**. What will change, however, are the technologies by which disinformation is created and spread.

The use of disinformation as a tool for influence and obfuscation will never cease. Similarly, the underlying psychological factors that make humans vulnerable to disinformation are enduring. What will change, however, are the technologies by which disinformation is created and spread. Indeed, this constantly changing arsenal of tools has led a majority of surveyed tech experts to conclude that the problem will not be solved within the next decade.[225]

Stakeholders face the distinct challenge of developing policy solutions to protect the information environment in a way that does not undermine public trust, while curbing a disinformation problem that will only continue evolving. What stakeholders should aim for, then, are strategies to mitigate disinformation and its potentially disastrous consequences while maintaining a robust commitment to civil liberties, freedom of expression, and privacy.

The proposed public-private partnership model included in this report is one possible approach for harnessing diverse expertise to solve this problem. Aligning industry and technical experts with the lawmakers who shape public policy will help produce an informed and measured response to a complex, rapidly transforming threat. It is to be expected that competing interests and incentives will hinder coordination, but a collaborative public-private framework is a prudent foundation on which to build consensus and coordinate action.

222. "Better Language Models and Their Implications," OpenAI, February 14, 2019, https://blog.openai.com/better-language-models/#sample2.

223. Julia Alexander, "Virtual creators aren't AI - but AI is coming for them," *The Verge*, 30 January 2019, https://www.theverge.com/2019/1/30/18200509/ai-virtual-creators-lil-miquela-instagram-artificial-intelligence.

224. Ingrid Angulo, "Facebook and YouTube should have learned from Microsoft's racist chatbot," CNBC, 17 March 2018, https://www.cnbc.com/2018/03/17/facebook-and-youtube-should-learn-from-microsoft-tay-racist-chatbot.html.

225 Janna Anderson and Lee Rainee, "The Future of Truth and Misinformation Online," Pew Research Center, 19 October 2017, http://www.pewinternetInternet.org/2017/10/19/the-future-of-truth-and-misinformation-online/.

WWW.PARK-ADVISORS.COM