B | Foreign Policy
at BROOKINGS

# THE ROLE OF TECHNOLOGY IN ONLINE MISINFORMATION

SARAH KREPS

JUNE 2020

## EXECUTIVE SUMMARY

States have long interfered in the domestic politics of other states. Foreign election interference is nothing new, nor are misinformation campaigns. The new feature of the 2016 election was the role of technology in personalizing and then amplifying the information to maximize the impact. As a 2019 Senate Select Committee on Intelligence report concluded, malicious actors will continue to weaponize information and develop increasingly sophisticated tools for personalizing, targeting, and scaling up the content.

This report focuses on those tools. It outlines the logic of digital personalization, which uses big data to analyze individual interests to determine the types of messages most likely to resonate with particular demographics. The report speaks to the role of artificial intelligence, machine learning, and neural networks in creating tools that distinguish quickly between objects, for example a stop sign versus a kite, or in a battlefield context, a combatant versus a civilian. Those same technologies can also operate in the service of misinformation through text prediction tools that receive user inputs and produce new text that is as credible as the original text itself. The report addresses potential policy solutions that can counter digital personalization, closing with a discussion of regulatory or normative tools that are less likely to be effective in countering the adverse effects of digital technology.

## INTRODUCTION

Meddling in domestic elections is nothing new as a tool of foreign influence. In the first two-party election in 1796, France engaged in aggressive propaganda[1] to tilt the public opinion scales in favor of the pro-French candidate, Thomas Jefferson, through a campaign of misinformation and fear.

The innovation of the 2016 presidential election, therefore, was not foreign interests or misinformation, but the technology used to promote those foreign interests and misinformation. Computational propaganda,[2] the use of big data and machine learning about user behavior to manipulate public opinion, allowed social media bots to target individuals or demographics known to be susceptible to politically sensitive messaging.

As the Senate Select Committee on Intelligence concluded,[3] the Russian Internet Research Agency (IRA) that used social media to divide and exercise American voters clearly understood American psychology and "what buttons to press." The IRA, however, did not fully exploit some of the technological tools that would have allowed it to achieve greater effect. In particular, the Senate report notes that the IRA did not use Facebook's

"Custom Audiences" feature that would have enabled more micro-targeting of advertisements on divisive issues. Nonetheless, Senator Mark Warner (D-VA) of the Intelligence Committee foreshadowed that the IRA and other malicious actors would remedy any previous shortcomings:

> There's no doubt that bad actors will continue to try to weaponize the scale and reach of social media platforms to erode public confidence and foster chaos. The Russian playbook is out in the open for other foreign and domestic adversaries to expand upon — and their techniques will only get more sophisticated.[4]

This report outlines the way that advances in digital technology will increasingly allow adversaries to expand their techniques in ways that drive misinformation. In particular, it speaks to the availability of user data and powerful artificial intelligence, a combination that enables platforms to personalize content. While conventional propaganda efforts were pitched to mass audiences and limited to manipulation of the median voter, tailored and personalized messages allow malicious actors to psychologically manipulate all corners of the ideological spectrum, thereby achieving a larger potential effect.

To make these points, the report first outlines the concept of digital personalization, in which users are targeted with content tailored to their interests and sensitivities. It then offers a discussion of how artificial intelligence fits into that digital personalization picture, integrating insights about machine learning and neural networks to show how algorithms can learn distinguish between objects or create synthetic text. The report next shows how AI can be used maliciously in the service of misinformation, focusing on text prediction tools that receive user inputs and produce styles and substance that are as credible as the original text itself. It then addresses potential policy solutions that can counter personalization via AI, and closes with a discussion of regulatory or normative tools that are less likely to be effective in countering the adverse effects of digital technology.

## PERSONALIZATION AT SCALE AND THE ROLE OF AI

In a November 2016 article, McKinsey Digital published an article[5] titled: "Marketing's Holy Grail: Digital personalization at scale." The authors note that an era of instant gratification means that customers decide quickly what they like, which means that companies must curate and deliver personalized content. "The tailoring of messages or offers based on their actual behavior" is key to luring and keeping consumers, the McKinsey article wrote. Step one in that journey is to understand consumer behavior with as much data as possible, which is where technology comes in. Big data combined with machine learning ensures that individuals receive "the appropriate trigger message," in the article's words.

Although pitched to companies, the personalization of content is not restricted to marketing. In 2016, the Russian IRA deployed similar principles in the 2016 election. According to the U.S. Senate Select Committee on Intelligence report "Russian Active Measures Campaigns and Interference in the 2016 Election," the IRA used targeted advertisements, falsified news articles, and social media amplification tools to polarize Americans.[6] Far from a strategy oriented toward a mass public, the IRA information operation relied on digital personalization: determining what types of sites individuals frequented, correlating between those behaviors and demographic information, and finding ways to reach the groups that would be most triggered by racially, ethnically, or religiously-charged content.

From there, the IRA could create thousands of microsegments, not all of which were created equal. In the election context, as the Senate Intelligence Committee report notes,[7] "no single group of Americans was targeted by IRA information operatives more than African-Americans." Social media content with racial undertones — whether advertisements, memes, or tweets — targeted African Americans, for example, with an eye toward generating resentment toward out-groups, co-opting participation in protest behavior, or even convincing individuals to sit out from elections. One advertisement[8] sought to stir up nativist sentiment through an image about Islam taking over the world, posted by an account called "Heart of Texas."

Micro-targeted messaging is not onerous, provided that large amounts of user data are available to generate profiles of personal likes, dislikes, ideology, and psychology. Generating new content that targets those micro-segments is, however, more resource-intensive.

Individuals who work at Russia's IRA work 12-hour shifts and are required[9] to meet quotas in terms of comments, blog posts, or page views. The work is tedious, and writing new content about the same topics or themes — for example, elevating the image of Russia or increasing division or confusion in the American political landscape — has its challenges. Enter artificial intelligence, which can help overcome these creativity obstacles.

## ADVANCES IN ARTIFICIAL INTELLIGENCE

Diving deeper into the ways that AI can facilitate misinformation requires taking a step back and examining the technology itself. The term "artificial intelligence" is one that is used frequently, but rarely uniformly. It refers generally to something the mathematician Alan Turing called a "thinking machine" that could process information like a human. In 1950, Turing wrote a paper called "Computing Machinery and Intelligence" that posed the question of whether machines can think.[10] He defined "think" as whether a computer can reason, the evidence being that humans would not be able to distinguish — in a blind test — between a human and a computer. Implied was that machines would be able to make judgments and reflections, and in intentional, purposive ways.[11]

Even if the differences between human and machine reasoning remain large, machines can think and indeed are increasingly outperforming humans on at least certain tasks. In 1997, IBM's chess-playing computer called Deep Blue beat the world chess champion Garry Kasparov. In 2015, AlphaGo, developed by DeepMind Technologies (later acquired by Google), defeated a human professional player of the game Go, considered more difficult for computers to win than chess because of the game structure.

Computers have gained these skills from advancements in artificial intelligence. Learning algorithms are generated through a process in which neural networks (a combination of neurons analogous to those in a human brain) make connections between cause and effect, or steps that are correct and incorrect. In the context of AlphaGo, the neural network analyzed millions of moves that human Go experts had made, then played against itself to reinforce what it had learned, fine-tuning and updating to predict and preempt moves.

Beyond the context of gaming, neural networks can classify images, video, or text by identifying patterns and shapes, engaging in logical reasoning about the identity of those images, and engaging in feedback corrections that improve the performance of the network. Training autonomous vehicle algorithms involves feeding the machine thousands of images and angles of stop signs, for example, so that the car can accurately recognize and heed a stop sign, even one that might be partially covered by a sticker, or so that the car does not stop for a kite that it mistakes for a stop sign.

Machine learning algorithms are enabling a number of technologies on similar principles of training the neural network with large amounts of data so that the machine can make connections, anticipate sequences of events, or classify new objects based on the resemblance with other objects. Utility companies collect large volumes of data on consumers' heating and air conditioning patterns to anticipate and regulate the flow of energy to households, notifying users of upcoming surges and encouraging behavior that increases efficiency across the grid, such as reducing consumption among some homes during peak hours.

In a defense policy context, a program called Project Maven was trained on terabytes of drone data to help differentiate people from objects. The project uses computer vision, a field of artificial intelligence that uses large amounts of digital images from videos combined with deep learning models to identify and classify objects. Instead of identifying the difference between a stop sign and a kite as in the example above — or a dog versus a cat, another common illustration of how neural networks learn to classify objects — the algorithm was trained to focus on 38 classes of objects that the military needed to detect on missions in the Middle East.[12] The military hastened to point out that the algorithm did not pick targets but provided faster and higher volume analysis than human analysts.[13]

As the 2016 election revealed, AI also offers potential tools of election interference through online misinformation, though not in a vacuum. Drawing on the 2016 Senate Intelligence Committee report, individuals seeking to meddle would start with an interdisciplinary study of social and political cleavages in a country, a demographic analysis of which groups occupy what space in those debates, and what types of content are most polarizing. They would then need a vehicle for efficiently generating that content, distributing and amplifying it, learning what was effective, and then repeating the process. The next section outlines specific AI tools that can generate misinformation at scale, aided by humans in the loop to determine the nature of division and content that might exploit those divisions.

## *Text prediction tools as a handmaiden to creativity and scale*

Research groups have begun developing text prediction tools that can produce fake or synthetic news at scale. One of the most prominent of these tools is called GPT-2, created by OpenAI, an independent research group whose stated aim is to promote ethical artificial intelligence. The model is trained on a dataset of 8 million web pages and predicts the next word or words on the basis of the previous words in the text that is offered as the input. OpenAI describes the model as "chameleon-like — it adapts to the style and content of the conditioning text," and without domain-specific training datasets.[14] Enter part of a poem, and the model will complete a poem by generating additional lines that match the style of the input. The same can be done with an Amazon product review, a greeting card, or — in the context of a plausible use case for misinformation — provocative text. To satisfy their writing quotas, an IRA employee could find text with suitable ideological or substantive perspective and enter it into the text prediction tool, which would produce unlimited numbers of articles that appear to be original and will not be identified as plagiarized.

A brief example illustrates the technology. The website ZeroHedge is technically a finance website, but it also engages in non-financial opinion writing that has been associated with alt-right views and has been banned periodically on Facebook and Twitter. In this case, the pro-Russia headline and first sentence acted as the conditioning text. The output below gives an example of one of the more refined, coherent GPT-2 outputs.

Conditioning text:

> Why Both Republicans And Democrats Want Russia To Become The Enemy Of Choice
>
> One of the more interesting aspects of the nauseating impeachment trial in the Senate was the repeated vilification of Russia and its President Vladimir Putin.

GPT-2 Output:

> No less than Senators from both parties described Putin as “a murderer” and an “emperor.” Democrats were so desperate for a reason to impeach Trump that they began calling him “Putin’s Pawn,” a “Kremlin Puppet” and a “Russian Agent.”

**The consequence is that citizens tune out of the political discourse or tune into their own, politically congenial filter bubble.**

The goal of mass-producing slanted information is not necessarily to change policy views, which is actually quite difficult. Rather, it is to confuse, polarize, and entrench. The more misinformation proliferates online, the more confused the average reader becomes, lost in a “fog of political news” as The New York Times concluded.[15] The consequence is that citizens tune out the political discourse or tune into their own, politically congenial filter bubble. A vicious cycle becomes nearly inevitable — people tune out perspectives that do not comport with their own, polarization becomes more entrenched, and midway compromise is nearly impossible. Since coherent policy requires shared reference points, the misinformation succeeds not by changing minds but by keeping people in their polarized lanes.

If the potential for misuse looms large, why have malicious actors not employed the technology to a greater extent? One possibility is that the technology is still relatively nascent. One of the most powerful versions of the GPT-2 was just released in November 2019. Far from flawless, it improved upon earlier versions that were far more likely to contain grammatical or factual errors, or simply be incoherent. For example, in one of the less powerful versions of GPT-2, conditioning text about North Korea from The New York Times (input by the author) produced the following gibberish:

> Life is a place full of surprises! Melt a glorious Easter cake in French but not that green. Well, a green cake, but for a Tuesday, of course! All Easter party year and here is the reason for baka.

The non-sensical content continued. Savvy actors could easily filter out this particular output and churn out more credible-sounding text. Advancements in the technology, however, have reduced the incoherent outputs and fostered more persuasive and credible text on average, which facilitates full automation by actors who seek to generate divisive online content. In a world where bots amplify content or only a few tweets need to be well-written and credible to go viral, then the imperfections in AI-generated text need not be deal-breakers. The systems may not be sophisticated enough to be used in entirely computationally-driven content creation without human oversight, but can be a useful vehicle for malicious actors who are looking to AI to overcome cognitive limitations and meet their content quotas.

Another possibility is that information is a form of currency, and the more it is deployed the less valuable it is. Take, for example, the initial deepfakes — which use deep learning to create manipulated media images or videos meant to look real[16] — of President Barack Obama, Speaker of the House Nancy Pelosi, or Facebook CEO Mark Zuckerberg, which carried enormous shock value. People quickly learned how to identify attributes of manipulated media, which rendered the deepfakes less powerful. Educational interventions, even those that are informal, are effective. Indeed, the scalability of text-generating models is a double-edged sword. On the one hand, it allows synthetic text to proliferate at the hands of one user. On the other hand, as the scale increases, consumers of online media also learn how to identify particular structures of sequences of text as fake, as in the case of deepfakes.[17] In the context of text, malicious actors might decide that rather than flooding the internet with synthetic text, they should deploy it more selectively in ways that would have more impact, such as before a major election.

Regardless of the fact that it has not yet been widely deployed, the ease and economical nature of the technology — as well as the effectiveness[18] in terms of producing text that readers deem to be as credible as human-generated text — raises the prospect of proliferation. AI text generation may be carried out in a fully automated way or, more likely, in conjunction with human curation. One set of use cases is benign and already here: sports box scores, stock quotes, product reviews. Another set of use cases may consist of misuse, as state and non-state actors find the tools to be a convenient way to generate convincing content. The question then is what to do about the less benign form of proliferation.

## POLICY SOLUTIONS

Although the technology for creating misinformation will only improve, so might the countermeasures. This section outlines the potential mechanisms through which particular public policy interventions might counter online misinformation.

### *Education interventions*

The most straightforward countermeasure is in some ways also the most difficult: public literacy interventions. A number of platforms have rolled out internet literacy initiatives to help users filter out misinformation online. Google has been running a digital safety[19] program to help young people identify fake news, including through understanding how artificial intelligence works, showing comparisons between chats with computer bots versus human beings, and identifying the markers of credible versus dubious sources or sites.

Similar logics hold for the more sophisticated AI-based misinformation campaigns. In the context of text prediction tools, certain features correspond to synthetic text, as highlighted above. As one study of fake news concluded, “people fall for fake news because they fail to think,” not because they fall prey to partisan or ideological bias.[20] Thinking, in the case of synthetic text, is looking both for the obvious grammatical or factual errors but also more subtle problems with the text.

To study the credibility of the text systematically, researchers generated text based on a New York Times story about North Korea.[21] One of the outputs cited “Ted Lieu (D-Calif), chairman of the Congressional Foreign Trade Committee.” Congressional committees are referred to as House or Senate, and no Foreign Trade Committee exists, let alone one on which he has a seat (he is on the House Foreign Affairs Committee). Moreover, states tend to be referred to by two-letter abbreviations rather than as “Calif.” Another story used the abbreviation DPRK and then followed with “DPRK is the initials of North Korean leader Kim Jong Un’s father,” which is inaccurate; it refers instead to the Democratic People’s Republic of Korea.

When readers think they are reading a news story, they are likely to take the facts at face value. Literacy campaigns would imply greater awareness about the prevalence of fake or synthetic news and trust in one’s own judgment, dismissing a story with dubious information rather than taking it as a given.

### *Technology as a response to tech-based misinformation*

One of the ways to resolve the problem of tech-based misinformation is through tech itself. The main mechanism for identifying inauthentic text is to use the same AI text generator. Since neural networks generate synthetic text, they are also familiar[22] with the text’s “habits, quirks, and traits” — which may also make them well-suited to detecting content that emerges from those networks. The Allen Institute for AI built a model named Grover, which not only generates neural fake news but also spots its own fake news and that of other AI generators. Studies of fake news detection found that it had a 92% accuracy in terms of detecting human- versus machine-written news.[23]

Relatedly, a collaboration between Harvard and MIT developed a forensic tool for detecting authenticity based on statistical probabilities about the likelihood that each word would be the predicted word. The tool is not the same AI text generator itself, but rather is an independent statistical analysis of text that embeds detection methods in a visual highlighting tool — somewhat like plagiarism software — which highlights text passages based on the model density of generated output compared to the human-generated text. In experimental tests with human subjects, the visual detection tool increased readers' ability to detect fake text from 54% to 72%.[24]

Another tech-based solution[25] involves analyzing metadata to identify synthetic text. Algorithms can be trained to identify the markers of malicious documents — such as the time written to produce the text, the number of accounts associated with a particular IP address, or the website itself — to identify malicious or inauthentic text. In responding to criticism about interference in the 2016 election, Facebook, for example, has used digital forensics and intelligence experts to identify posts and accounts from either inauthentic users. All of these posts were meant to polarize the target users, largely in North Africa, Latin America, and the United States.[26]

Independent users have begun programming tools that implement the underlying approaches to identifying fake text. One tool consists of an extension for the internet browser Chrome that detect text written by neural nets, comparing the output to GPT-2's own recommendations. However, the tool has flaws. It generates a number of false positives: For example, the tool gave a low likelihood that excerpts from James Joyce's "Ulysses" and a Donald Trump speech were real. Further, tweaking the neural network on which the tool is based would foil the extension and render it ineffective. Malicious actors looking to spread misinformation and those trying to counter it are involved in a cat-and-mouse game, in which counter-measures lead to modifications of the original approach and inevitable challenges arise in addressing the source of misinformation. The challenge reflects that of detecting synthetics more generally, whether they are deepfake videos, text, or imagery. As research produces advances in detection, whether for individuals to cue on the attributes of fakeness or technology to facilitate that detection, the synthetics themselves become more sophisticated, making any advances ephemeral.[27]

## CONCLUSION

As this report suggests, the incentives toward personalization in the commercial sector and advancements in AI that accelerate personalization combine to create vulnerabilities in the form of online misinformation. AI can now create credible information at scale, overcoming the limitations of human capacity to produce misinformation. Then, based on studies of social and political cleavages in society, malicious actors can target particular content at groups that would be most susceptible to certain divisive messages. Although different policy interventions — including education or digital literacy and technology itself — might mitigate the vulnerabilities, personalization via AI remains a powerful force, with data at its root. Since 2016, every social media platform has taken aggressive measures to protect users' privacy, and governments such as the European Union (with its General Data Protection Regulation, or GDPR) have developed policies aimed at data protection and privacy. To be sure, social media sites can still be hacked and harvested for data, but the near-ubiquitous awareness of privacy settings and the sites' awareness that profitability hinges on user trust would suggest that valuable steps have been and can still be taken to address the data privacy issues that might be associated with personalization.

**Understanding the potential misuse cases is more practical than trying to contain a technology whose underlying AI fundamentals are fairly well understood.**

To the extent that text generation for developing misinformation at scale creates opportunities for foreign election interference or influence of another country's domestic politics more generally, then the question is whether these tools should be legally or normatively proscribed. Groups such as OpenAI have experimented with timed and deliberate releases of these tools, studying in advance the potential for misuse, collaborating with researchers to understand the potential consequences, and releasing when it does not see evidence of misuse.[28] Although critics suggest that the tool can enable maliciousness and the staged release can produce hysteria about AI, a convincing counterargument suggests that the technological cat is already out of the bag.[29] Understanding the potential misuse cases is more practical than trying to contain a technology whose underlying AI fundamentals are fairly well understood.

Similarly, regulatory moves may prove challenging. The proposed Digital Services Act in the European Union, which would regulate online platforms in the EU, could consider proscribing text prediction tools, except that analogous tools are already ubiquitous in non-malicious contexts and would therefore create a number of false positives for any AI text detection tools. Box scores, stock market summaries, and earthquake alerts are just some of the many applications of text prediction tools. Even provided that technology can identify synthetic text, almost all of the hits would be non-malicious applications, meaning that any regulatory move to prohibit the use of these tools could flag a lot of benign content. In the United States, Section 230 of the Communications Decency Act, which offers protection for blocking or screening offensive material, would have similar challenges: Identifying the offensive material such that it does not violate free speech requirements would be difficult because of the likelihood of false positives.

More fruitful is greater individual awareness of the proliferation of personalization AI and of malicious actors' temptation to make use of these tools. As Special Counsel Robert Mueller testified in 2019 regarding Russian election interference: "They're doing it as we sit here."[30] They may be engaging in influence operations through the combination of personalization and AI-generated content. It behooves online consumers to be aware of and guard against this threat.

# REFERENCES

1 Alden Fletcher, "Foreign Election Interference in the Founding Era," Lawfare, October 25, 2018, https://www.lawfareblog.com/foreign-election-interference-founding-era.

2 "The Computational Propaganda Project," Oxford Internet Institute, https://comprop.oii.ox.ac.uk/.

3 "Report of The Select Committee on Intelligence of the United States Senate on Russian Active Measures Campaigns and Interference in the 2016 U.S. Election: Volume 2: Russia's Use of Social Media with Additional Views," (Washington, DC: United States Senate, October 8, 2019), https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume2.pdf, p.41.

4 "Senate Intel Committee Releases Bipartisan Report on Russia's Use of Social Media," Office of Senator Richard Burr, October 10, 2019, https://www.burr.senate.gov/press/releases/senate-intel-committee-releases-bipartisan-report-on-russias-use-of-social-media-.

5 Brian Gregg, "Marketing's Holy Grail: Digital Personalization at scale," McKinsey, November 18, 2016, https://www.mckinsey.com/business-functions/mckinsey-digital/our-insights/marketings-holy-grail-digital-personalization-at-scale.

6 "Report of The Select Committee on Intelligence of the United States Senate on Russian Active Measures Campaigns and Interference in the 2016 US Election, Vol 2," United States Senate.

7 Ibid.

8 Nitasha Tiku, "How Russia 'Pushed Our Buttons' With Fake Online Ads," *Wired*, November 3, 2017, https://www.wired.com/story/how-russia-pushed-our-buttons-with-fake-online-ads/.

9 Neil MacFarquhar, "Inside the Russian Troll Factory: Zombies and a Breakneck Pace," *The New York Times*, February 18, 2018, https://www.nytimes.com/2018/02/18/world/europe/russia-troll-factory.html.

10 AM Turing, "Computing Machinery and intelligence," *Mind* 59, no. 236 (October 1950): 433-460, https://academic.oup.com/mind/article/LIX/236/433/986238.

11 Darrell West, "What is artificial intelligence," The Brookings Institution, October 4, 2018, https://www.brookings.edu/research/what-is-artificial-intelligence/.

12 Cheryl Pellerin, "Project Maven to Deploy Computer Algorithms to War Zone by Year's End," DOD News, July 21, 2017.

13 Kelsey Atherton, "Targeting the future of the DoD's controversial Project Maven initiative," *C4ISRNet*, 27 July 2018.

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

15 Sabrina Tavernise and Aidan Gardiner, "'No One Believes Anything': Voters Worn Out by a. Fog of Political News," *The New York Times*, November 18, 2019, https://www.nytimes.com/2019/11/18/us/polls-media-fake-news.html.

16 Grace Shao, "What 'deepfakes' are and how they may be dangerous," CNBC, 13 October 2019.

17 Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi, "Defending Against Neural Fake News," (Ithaca, NY: Cornell University, May 2019): 10-35, https://arxiv.org/abs/1905.12616.

18 Sarah E. Kreps, Miles McCain, and Miles Brundage, "All the News that's Fit to Fabricate: AI-Generated Text as a Tool of Media Misinformation," (SSRN, January 2020), https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3525002.

19 Sarah Perez, "Google's new media literacy program teaches kids how to spot disinformation and fake news," TechCrunch, June 24, 2019, https://techcrunch.com/2019/06/24/googles-new-media-literacy-program-teaches-kids-how-to-spot-disinformation-and-fake-news/.

20 Gordon Pennycook and David Rand, "Lazy, Not Biased: Susceptibility to Partisan Fake News Is Better Explained by Lack of Reasoning Than by Motivated Reasoning," *Cognition* 188 (July 2019): 39-50, https://doi.org/10.1016/j.cognition.2018.06.011.

21 Sarah Kreps, Miles McCain, and Miles Brundage, "All the News that's Fit to Fabricate."

22 "GROVER — A State-of-the-Art Defense against Neural Fake News," Allen Institute for AI, https://grover.allenai.org/detect.

23 Rowan Zellers, Ari Holtzman, Hannah Rashkin, Yonatan Bisk, Ali Farhadi, Franziska Roesner, and Yejin Choi, "Defending against Neural Fake News."

24 Sebastian Gehrmann, Hendrik Strobelt, and Alexander Rush, "GLTR: Statistical Detection and Visualization of Generated Text," (Ithaca, NY: Cornell University, June 2019), https://arxiv.org/abs/1906.04043.

25 Irene Solaiman, Miles Brundage, Jack Clark, Amanda Askell, Ariel Herbert-Voss, Jeff Wu, Alec Radford, Gretchen Krueger, Jong Wook Kim, Sarah Kreps, Miles McCain, Alex Newhouse, Jason Blazakis, Kris McGuffie, and Jasmine Wang, "Release Strategies and the Social Impacts of Language Models," (Ithaca, NY: Cornell University, November 2019), https://arxiv.org/abs/1908.09203.

26 Mike Isaac, "Facebook finds new disinformation campaigns and braces for 2020 torrent," *The New York Times*, October 21, 2019, https://www.nytimes.com/2019/10/21/technology/facebook-disinformation-russia-iran.html.

27 Alex Engler, "Fighting deepfakes when detection fails," (Washington, DC: The Brookings Institution, November 14, 2019), https://www.brookings.edu/research/fighting-deepfakes-when-detection-fails/.

28 Karen Hao, "OpenAI has released the largest version yet of its fake-news-spewing AI," *MIT Technology Review*, August 29, 2019, https://www.technologyreview.com/2019/08/29/133218/openai-released-its-fake-news-ai-gpt-2/.

29 James Vincent, "AI researchers debate the ethics of sharing potentially harmful programs," The Verge, February 21, 2019, https://www.theverge.com/2019/2/21/18234500/ai-ethics-debate-researchers-harmful-programs-openai.

30 Philip Ewing, "Mueller On Russian Election Interference: 'They're Doing It As We Sit Here,'" NPR, July 24, 2019, https://www.npr.org/2019/07/24/743093777/watch-live-mueller-testifies-on-capitol-hill-about-2016-election-interference.

## ABOUT THE AUTHOR

**Sarah Kreps** is the John L. Wetherill Professor of Government, Milstein Faculty Fellow in Technology and Humanity, and an adjunct professor of law at Cornell University. She is the author of five books, including most recently, *Social Media and International Politics* (Cambridge University Press, forthcoming). Between 1999 and 2003, she served as an active duty officer in the United States Air Force. She has a bachelor's degree from Harvard University, Master of Science degree from the University of Oxford, and doctorate from Georgetown University.

## ACKNOWLEDGEMENTS

The author would like to thank Elliott Ash, Miles Brundage, Miles McCain, and Chris Meserole for their comments on this research. The author is also grateful for the valuable feedback of seminar participants at Oxford's Centre for the Governance of AI, Cornell Tech's Digital Life Initiative, ETH Zurich's Center for Law and Economics, and the Center for New American Security's AI and International Security workshop. Anna Newby edited this policy brief, with assistance from Caroline Klaff, and Rachel Slattery performed layout.

The Brookings Institution is a nonprofit organization devoted to independent research and policy solutions. Its mission is to conduct high-quality, independent research and, based on that research, to provide innovative, practical recommendations for policymakers and the public. The conclusions and recommendations of any Brookings publication are solely those of its author(s), and do not reflect the views of the Institution, its management, or its other scholars.