



COUNTERING DISINFORMATION EFFECTIVELY

An Evidence-Based Policy Guide

Jon Bateman and Dean Jackson



CARNEGIE
ENDOWMENT FOR
INTERNATIONAL PEACE

COUNTERING DISINFORMATION EFFECTIVELY

An Evidence-Based Policy Guide

Jon Bateman and Dean Jackson

This research was supported by a grant from the Special Competitive Studies Project.

© 2024 Carnegie Endowment for International Peace. All rights reserved.

Carnegie does not take institutional positions on public policy issues; the views represented herein are those of the author(s) and do not necessarily reflect the views of Carnegie, its staff, or its trustees.

No part of this publication may be reproduced or transmitted in any form or by any means without permission in writing from the Carnegie Endowment for International Peace. Please direct inquiries to:

Carnegie Endowment for International Peace
Publications Department
1779 Massachusetts Avenue NW
Washington, DC 20036
P: + 1 202 483 7600
F: + 1 202 483 1840
CarnegieEndowment.org

This publication can be downloaded at no cost at CarnegieEndowment.org.

TABLE OF CONTENTS

About the Authors	v
Acknowledgments	vii
Summary	1
Methodology	9
Challenges and Cautions	11
CASE STUDY 1 Supporting Local Journalism	17
CASE STUDY 2 Media Literacy Education	23
CASE STUDY 3 Fact-Checking	29

CASE STUDY 4 Labeling Social Media Content	35
CASE STUDY 5 Counter-messaging Strategies	43
CASE STUDY 6 Cybersecurity for Elections and Campaigns	51
CASE STUDY 7 Statecraft, Deterrence, and Disruption	57
CASE STUDY 8 Removing Inauthentic Asset Networks	63
CASE STUDY 9 Reducing Data Collection and Targeted Ads	71
CASE STUDY 10 Changing Recommendation Algorithms	79
Looking Ahead: Generative AI	85
Notes	89
Carnegie Endowment for International Peace	119

ABOUT THE AUTHORS

JON BATEMAN is a senior fellow in the Technology and International Affairs Program at the Carnegie Endowment for International Peace. His research areas include disinformation, cyber operations, artificial intelligence, and techno-nationalism. Bateman previously was special assistant to Chairman of the Joint Chiefs of Staff General Joseph F. Dunford, Jr., serving as a speechwriter and the lead strategic analyst in the chairman's internal think tank. He has also helped craft policy for military cyber operations in the Office of the Secretary of Defense, and was a senior intelligence analyst at the Defense Intelligence Agency, where he led teams responsible for assessing Iran's internal stability, senior-level decisionmaking, and cyber activities. Bateman is a graduate of Harvard Law School and Johns Hopkins University.

DEAN JACKSON is principal of Public Circle Research & Consulting and a specialist in democracy, media, and technology. In 2023, he was named an inaugural Tech Policy Press reporting fellow and an affiliate fellow with the Propaganda Research Lab at the University of Texas at Austin. Previously, he was an investigative analyst with the Select Committee to Investigate the January 6th Attack on the U.S. Capitol and project manager of the Influence Operations Researchers' Guild at the Carnegie Endowment for International Peace. From 2013 to 2021, Jackson managed research and program coordination activities related to media and technology at the National Endowment for Democracy. He holds an MA in international relations from the University of Chicago and a BA in political science from Wright State University in Dayton, OH.

ACKNOWLEDGMENTS

The authors wish to thank William Adler, Dan Baer, Albin Birger, Kelly Born, Jessica Brandt, David Broniatowski, Monica Bulger, Ciaran Cartmell, Mike Caulfield, Tímea Červeňová, Rama Elluru, Steven Feldstein, Beth Goldberg, Stephanie Hankey, Justin Hendrix, Vishnu Kannan, Jennifer Kavanagh, Rachel Kleinfeld, Samantha Lai, Laura Livingston, Peter Mattis, Tamar Mitts, Brendan Nyhan, George Perkovich, Martin Riedl, Ronald Robertson, Emily Roseman, Jen Rosiere Reynolds, Zeve Sanderson, Bret Schafer, Leah Selig Chauhan, Laura Smillie, Rory Smith, Victoria Smith, Kate Starbird, Josh Stearns, Gerald Torres, Meaghan Waff, Alicia Wanless, Laura Waters, Gavin Wilde, Kamy Yadav, and others for their valuable feedback and insights. Additional thanks to Joshua Sullivan for research assistance and to Alie Brase, Lindsay Maizland, Anjuli Das, Jocelyn Soly, Amy Mellon, and Jessica Katz for publications support. The final report reflects the views of the authors only. This research was supported by a grant from the Special Competitive Studies Project.

SUMMARY

Disinformation is widely seen as a pressing challenge for democracies worldwide. Many policymakers are grasping for quick, effective ways to dissuade people from adopting and spreading false beliefs that degrade democratic discourse and can inspire violent or dangerous actions. Yet disinformation has proven difficult to define, understand, and measure, let alone address.

Even when leaders know what they want to achieve in countering disinformation, they struggle to make an impact and often don't realize how little is known about the effectiveness of policies commonly recommended by experts. Policymakers also sometimes fixate on a few pieces of the disinformation puzzle—including novel technologies like social media and artificial intelligence (AI)—without considering the full range of possible responses in realms such as education, journalism, and political institutions.

This report offers a high-level, evidence-informed guide to some of the major proposals for how democratic governments, platforms, and others can counter disinformation. It distills core insights from empirical research and real-world data on ten diverse kinds of policy interventions, including fact-checking, foreign sanctions, algorithmic adjustments, and counter-messaging campaigns. For each case study, we aim to give policymakers an informed sense of the prospects for success—bridging the gap between the mostly meager scientific understanding and the perceived need to act. This means answering three core questions: How much is known about an intervention? How effective does the intervention seem, given current knowledge? And how easy is it to implement at scale?

OVERALL FINDINGS

- **There is no silver bullet or “best” policy option.** None of the interventions considered in this report were simultaneously well-studied, very effective, and easy to scale. Rather, the utility of most interventions seems quite uncertain and likely depends on myriad factors that researchers have barely begun to probe. For example, the precise wording and presentation of social media labels and fact-checks can matter a lot, while counter-messaging campaigns depend on a delicate match of receptive audiences with credible speakers. Bold claims that any one policy is the singular, urgent solution to disinformation should be treated with caution.
- **Policymakers should set realistic expectations.** Disinformation is a chronic historical phenomenon with deep roots in complex social, political, and economic structures. It can be seen as jointly driven by forces of supply and demand. On the supply side, there are powerful political and commercial incentives for some actors to engage in, encourage, or tolerate deception, while on the demand side, psychological needs often draw people into believing false narratives. Credible options exist to curb both supply and demand, but technocratic solutionism still has serious limits against disinformation. Finite resources, knowledge, political will, legal authority, and civic trust constrain what is possible, at least in the near- to medium-term.
- **Democracies should adopt a portfolio approach to manage uncertainty.** Policymakers should act like investors, pursuing a diversified mixture of counter-disinformation efforts while learning and rebalancing over time. A healthy policy portfolio would include tactical actions that appear well-researched or effective (like fact-checking and labeling social media content). But it would also involve costlier, longer-term bets on promising structural reforms (like supporting local journalism and media literacy). Each policy should come with a concrete plan for ongoing reassessment.
- **Long-term, structural reforms deserve more attention.** Although many different counter-disinformation policies are being implemented in democracies, outsized attention goes to the most tangible, immediate, and visible actions. For example, platforms, governments, and researchers routinely make headlines for announcing the discovery or disruption of foreign and other inauthentic online networks. Yet such actions, while helpful, usually have narrow impacts. In comparison, more ambitious but slower-moving efforts to revive local journalism and improve media literacy (among other possibilities) receive less notice despite encouraging research on their prospects.

- **Platforms and tech cannot be the sole focus.** Research suggests that social media platforms help to fuel disinformation in various ways—for example, through recommendation algorithms that encourage and amplify misleading content. Yet digital platforms exist alongside, and interact with, many other online and offline forces. The rhetoric of political elites, programming on traditional media sources like TV, and narratives circulating among trusted community members are all highly influential in shaping people’s speech, beliefs, and behaviors. At the same time, the growing number of digital platforms dilutes the effectiveness of actions by any single company to counter disinformation. Given this interplay of many voices and amplifiers, effective policy will involve complementary actions in multiple spheres.
- **Countering disinformation is not always apolitical.** Those working to reduce the spread and impact of disinformation often see themselves as disinterested experts and technocrats—operating above the fray of political debate, neither seeking nor exercising political power. Indeed, activities like removing inauthentic social media assets are more or less politically neutral. But other efforts, such as counter-messaging campaigns that use storytelling or emotional appeals to compete with false ideas at a narrative and psychological level, can be hard to distinguish from traditional political advocacy. Ultimately, any institutional effort to declare what is true and what is false—and to back such declarations with power, resources, or prestige—implies some claim of authority and therefore can be seen as having political meaning (and consequences). Denying this reality risks encouraging overreach, or inviting blowback, which deepens distrust.
- **Research gaps are pervasive.** The relatively robust study of fact-checking offers clues about the possibilities and the limits of future research on other countermeasures. On the one hand, dedicated effort has enabled researchers to validate fact-checking as a generally useful tool. Policymakers can have some confidence that fact-checking is worthy of investment. On the other hand, researchers have learned that fact-checking’s efficacy can vary a lot depending on a host of highly contextual, poorly understood factors. Moreover, numerous knowledge gaps and methodological biases remain even after hundreds of published studies on fact-checking. Because fact-checking represents the high-water mark of current knowledge about counter-disinformation measures, it can be expected that other measures will likewise require sustained research over long periods—from fundamental theory to highly applied studies.

- **Research is a generational task with uncertain outcomes.** The knowledge gaps highlighted in this report can serve as a road map for future research. Filling these gaps will take more than commissioning individual studies; major investments in foundational research infrastructure, such as human capital, data access, and technology, are needed. That said, social science progresses slowly, and it rarely yields definite answers to the most vexing current questions. Take economics, for example: a hundred years of research has helped Western policymakers curb (though not eliminate) depressions, recessions, and panics—yet economists still debate great questions of taxes and trade and are reckoning only belatedly with catastrophic climate risks. The mixed record of economics offers a sobering benchmark for the study of disinformation, which is a far less mature and robust field.
- **Generative AI will have complex effects but might not be a game changer.** Rapid AI advances could soon make it much easier and cheaper to create realistic and/or personalized false content. Even so, the net impact on society remains unclear. Studies suggest that people’s willingness to believe false (or true) information is often not primarily driven by the content’s level of realism. Rather, other factors such as repetition, narrative appeal, perceived authority, group identification, and the viewer’s state of mind can matter more. Meanwhile, studies of microtargeted ads—already highly data-driven and automated—cast doubt on the notion that personalized messages are uniquely compelling. Generative AI can also be used to counter disinformation, not just foment it. For example, well-designed and human-supervised AI systems may help fact-checkers work more quickly. While the long-term impact of generative AI remains unknown, it’s clear that disinformation is a complex psychosocial phenomenon and is rarely reducible to any one technology.

CASE STUDY SUMMARIES

1. **Supporting Local Journalism.** There is strong evidence that the decline of local news outlets, particularly newspapers, has eroded civic engagement, knowledge, and trust—helping disinformation to proliferate. Bolstering local journalism could plausibly help to arrest or reverse such trends, but this has not been directly tested. Cost is a major challenge, given the expense of quality journalism and the depth of the industry’s financial decline. Philanthropy can provide targeted support, such as seed money for experimentation. But a long-term solution would probably require government intervention and/or alternate business models. This could include direct subsidies (channeled through nongovernmental intermediaries) or indirect measures, such as tax exemptions and bargaining rights.

Table 1. Overview of Case Studies¹

Type	Intervention	How much is known?	How effective does it seem?	How easily does it scale?
	1. Supporting local journalism	Modest	Significant	Difficult
	2. Media literacy education	Significant	Significant	Difficult
	3. Fact-checking	Significant	Modest	Modest
	4. Labeling social media content	Modest	Modest	Easy
	5. Counter-messaging strategies	Modest	Modest	Difficult
	6. Cybersecurity for elections and campaigns	Modest	Modest	Modest
	7. Statecraft, deterrence, and disruption	Modest	Limited	Modest
	8. Removing inauthentic asset networks	Limited	Modest	Modest
	9. Reducing data collection and targeted ads	Modest	Limited	Difficult
	10. Changing recommendation algorithms	Limited	Significant	Modest
 Public information	 Government action	 Platform action		

2. **Media Literacy Education.** There is significant evidence that media literacy training can help people identify false stories and unreliable news sources. However, variation in pedagogical approaches means the effectiveness of one program does not necessarily imply the effectiveness of another. The most successful variants empower motivated individuals to take control of their media consumption and seek out high-quality information—instilling confidence and a sense of responsibility alongside skills development. While media literacy training shows promise, it suffers challenges in speed, scale, and targeting. Reaching large numbers of people, including those most susceptible to disinformation, is expensive and takes many years.
3. **Fact-Checking.** A large body of research indicates that fact-checking can be an effective way to correct false beliefs about specific claims, especially for audiences that are not heavily invested in the partisan elements of the claims. However, influencing factual beliefs does not necessarily result in attitudinal or behavioral changes, such as reduced support for a deceitful politician or a baseless policy proposal. Moreover, the efficacy of fact-checking depends a great deal on contextual factors—such as wording, presentation, and source—that are not well understood. Even so, fact-checking seems unlikely to cause a backfire effect that leads people to double down on false beliefs. Fact-checkers face a structural disadvantage in that false claims can be created more cheaply and disseminated more quickly than corrective information; conceivably, technological innovations could help shift this balance.
4. **Labeling Social Media Content.** There is a good body of evidence that labeling false or untrustworthy content with additional context can make users less likely to believe and share it. Large, assertive, and disruptive labels are the most effective, while cautious and generic labels often do not work. Reminders that nudge users to consider accuracy before resharing show promise, as do efforts to label news outlets with credibility scores. Different audiences may react differently to labels, and there are risks that remain poorly understood: labels can sometimes cause users to become either overly credulous or overly skeptical of unlabeled content, for example. Major social media platforms have embraced labels to a large degree, but further scale-up may require better information-sharing or new technologies that combine human judgment with algorithmic efficiency.
5. **Counter-messaging Strategies.** There is strong evidence that truthful communications campaigns designed to engage people on a narrative and psychological level are more effective than facts alone. By targeting the deeper feelings and ideas that make false claims appealing, counter-messaging strategies have the potential to impact harder-to-reach audiences. Yet success depends on the complex interplay of many inscrutable factors. The best campaigns use careful audience analysis to select

the most resonant messengers, mediums, themes, and styles—but this is a costly process whose success is hard to measure. Promising techniques include communicating respect and empathy, appealing to prosocial values, and giving the audience a sense of agency.

6. **Cybersecurity for Elections and Campaigns.** There is good reason to think that campaign- and election-related cybersecurity can be significantly improved, which would prevent some hack-and-leak operations and fear-inducing breaches of election systems. The cybersecurity field has come to a strong consensus on certain basic practices, many of which remain unimplemented by campaigns and election administrators. Better cybersecurity would be particularly helpful in preventing hack-and-leaks, though candidates will struggle to prioritize cybersecurity given the practical imperatives of campaigning. Election systems themselves can be made substantially more secure at a reasonable cost. However, there is still no guarantee that the public would perceive such systems as secure in the face of rhetorical attacks by losing candidates.
7. **Statecraft, Deterrence, and Disruption.** Cyber operations targeting foreign influence actors can temporarily frustrate specific foreign operations during sensitive periods, such as elections, but any long-term effect is likely marginal. There is little evidence to show that cyber operations, sanctions, or indictments have achieved strategic deterrence, though some foreign individuals and contract firms may be partially deterrable. Bans on foreign platforms and state media outlets have strong first-order effects (reducing access to them); their second-order consequences include retaliation against democratic media by the targeted state. All in all, the most potent tool of statecraft may be national leaders' preemptive efforts to educate the public. Yet in democracies around the world, domestic disinformation is far more prolific and influential than foreign influence operations.
8. **Removing Inauthentic Asset Networks.** The detection and removal from platforms of accounts or pages that misrepresent themselves has obvious merit, but its effectiveness is difficult to assess. Fragmentary data—such as unverified company statements, draft platform studies, and U.S. intelligence—suggest that continuous takedowns might be capable of reducing the influence of inauthentic networks and imposing some costs on perpetrators. However, few platforms even claim to have achieved this, and the investments required are considerable. Meanwhile, the threat posed by inauthentic asset networks remains unclear: a handful of empirical studies suggest that such networks, and social media influence operations more generally, may not be very effective at spreading disinformation. These early findings imply that platform takedowns may receive undue attention in public and policymaking discourse.

- 9. Reducing Data Collection and Targeted Ads.** Data privacy protections can be used to reduce the impact of microtargeting, or data-driven personalized messages, as a tool of disinformation. However, nascent scholarship suggests that microtargeting—while modestly effective in political persuasion—falls far short of the manipulative powers often ascribed to it. To the extent that microtargeting works, privacy protections seem to measurably undercut its effectiveness. But this carries high economic costs—not only for tech and ad companies, but also for small and medium businesses that rely on digital advertising. Additionally, efforts to blunt microtargeting can raise the costs of political activity in general, especially for activists and minority groups who lack access to other communication channels.
- 10. Changing Recommendation Algorithms.** Although platforms are neither the sole sources of disinformation nor the main causes of political polarization, there is strong evidence that social media algorithms intensify and entrench these off-platform dynamics. Algorithmic changes therefore have the potential to ameliorate the problem; however, this has not been directly studied by independent researchers, and the market viability of such changes is uncertain. Major platforms’ optimizing for something other than engagement would undercut the core business model that enabled them to reach their current size. Users could opt in to healthier algorithms via middleware or civically minded alternative platforms, but most people probably would not. Additionally, algorithms are blunt and opaque tools: using them to curb disinformation would also suppress some legitimate content.

METHODOLOGY

This report offers high-level, evidence-informed assessments of ten commonly proposed ways to counter disinformation. It summarizes the quantity and quality of research, the evidence of efficacy, and the ease of scalable implementation. Building on other work that has compiled policy proposals or collected academic literature, this report seeks to synthesize social science and practitioner knowledge for an audience of policymakers, funders, journalists, and others in democratic countries.² Rather than recommending a specific policy agenda, it aims to clarify key considerations that leaders should weigh based on their national and institutional contexts, available resources, priorities, and risk tolerance.

To conduct this research, we compiled a list of nearly two dozen counter-disinformation measures frequently proposed by experts, scholars, and policymakers.³ We then selected ten for inclusion based on several factors. First, we prioritized proposals that had a fairly direct connection to the problem of disinformation. For example, we excluded antitrust enforcement against tech companies because it affects disinformation in an indirect way, making it difficult to evaluate in this report. Second, we focused on countermeasures that could plausibly be subject to meaningful empirical study. We therefore did not consider diplomatic efforts to build international norms against disinformation, for example, or changes to platforms' legal liability as intermediaries. Third, we sought to cover a diverse range of interventions. This meant including actions implementable by the government, the private sector, and civil society; tactical measures as well as structural reforms; and multiple theories of change such as resilience, disruption, and deterrence.

The ten selected interventions became the subjects of this report's ten case studies. Each case study defines the intervention, gives concrete use cases, and highlights additional reading. The case studies focus on three questions: How much is known about an intervention? How effective does it seem, given current knowledge? And how easy is it to implement at scale? To develop these case studies, we reviewed hundreds of academic papers, previous meta-analyses, programmatic literature, and other relevant materials. We also conducted a series of workshops and consultations with scholars, practitioners, policymakers, and funders. We drew on experts with domain knowledge to vet individual case studies, as well as those with a broader view of the counter-disinformation field to provide feedback on the project as a whole. The resulting report expresses the views of the authors alone.

Although this report reviews a number of important, commonly proposed policy ideas, it is not comprehensive. In particular, we did not study the following significant categories of long-term, large-scale change. First, political institutions could try to perform stronger gate-keeping functions. This may involve reforms of party primaries, redistricting processes, and campaign finance systems. Second, tech platforms might need stronger incentives and capacity to curb disinformation. This could involve new regulation, diversification of revenue, and market power reductions that enable users, advertisers, activists, and others to provide checks on major platforms. Third, the public may need more encouragement to value truth and place trust in truthful institutions and figures. This might involve addressing the many root causes of popular alienation, fear, and anger, such as with local community-building efforts, a reversal of geographic sorting, improvements to economic prospects, and healing of racial grievances. Any of these ideas would be daunting to implement, and none are easy to assess. But they all have serious potential to help counter disinformation—perhaps even more so than the ten interventions studied in this report.

CHALLENGES AND CAUTIONS

Before seeking to counter disinformation, policymakers should carefully consider what this idea means. “Disinformation,” usually defined as information known by the speaker to be false, is a notoriously tricky concept that comes with numerous limitations, contradictions, and risks.⁴

CONCEPTUAL CHALLENGES

Identifying disinformation presents several puzzles. For one thing, labeling any claim as false requires invoking an authoritative truth. Yet the institutions and professions most capable of discerning the truth—such as science, journalism, and courts—are sometimes wrong and often distrusted. Moreover, true facts can be selectively assembled to create an overall narrative that is arguably misleading but not necessarily false in an objective sense. This may be even more common and influential than outright lies, yet it’s unclear whether it counts as disinformation. In fact, “disinformation” is frequently conflated with a range of other political and societal maladies such as polarization, extremism, and hate. All of these are technically distinct issues, though they can be causally related to disinformation and to each other. Finally, it is difficult to know whether someone spreading false claims does so intentionally. Disinformation typically passes through a long chain of both witting and unwitting speakers.

The challenges of the term “disinformation” are not merely theoretical; they have influenced public debates. Despite the word’s scientific-sounding imprimatur, it is often invoked quite loosely to denigrate any viewpoint seen as wrong, baseless, disingenuous, or harmful. Such usage has the effect of pathologizing swaths of routine discourse: after all, disagreements about what is wrong, baseless, disingenuous, or harmful are what drives democratic politics and social change. Moreover, today’s talk of “disinformation” can sometimes imply a more novel, solvable problem than really exists. Although the word has been familiar in the West for decades, it attained new currency just a few years ago after a series of catalyzing episodes—such as Russian election interference in the United States—involving social media. This led many people to see social media as the defining cause of disinformation, rather than one driver or manifestation of it. The messy battle for truth is, of course, an eternal aspect of human society.

For policymaking, reliance on a loaded but vague idea like “disinformation” brings several risks. When the term is used to imply that normal and necessary public discourse is dangerously disordered, it encourages the empowerment of technocrats to manage speech and, in turn, potentially erodes legal and normative boundaries that sustain democracy. Moreover, the term’s vagaries and contradictions are already well understood by segments of the public and have been seized upon, including by disinformers themselves, to undermine counter-disinformation efforts. In some cases, those accused of spreading disinformation have successfully sought to reclaim the term by arguing that counter-disinformation efforts are the real sources of disinformation, thus reversing the roles of perpetrator and victim.

This risk is most obvious in authoritarian regimes and flawed democracies, where leaders may suppress dissent by labeling it disinformation. But the problem can manifest in other ways too. A prominent U.S. example was the 2020 public letter by former intelligence officials warning that the then-recent disclosure of Hunter Biden’s laptop data “has all the classic earmarks of a Russian information operation.”⁵ Later, when the data’s authenticity was largely confirmed, those promoting the laptop story said the letter itself was a form of disinformation.⁶ Similar boomerang patterns have previously been seen with “fake news,” a phrase that originally described unethical content farms but was quickly repurposed to delegitimize truthful journalism. To be sure, such boomerangs often rest on exaggerated or bad faith claims. Yet they exploit a core truth: “disinformation” is a flawed, malleable term whose implied assertion of authority can lead to overreach and blowback.

For these and other reasons, a growing number of experts reject the term “disinformation.” Some prefer to focus instead on “misinformation” (which elides intent) or “influence/information operations” (which de-emphasizes falsity). Others favor more self-consciously political terms such as “propaganda” or “information warfare,” which they see as clearer warnings of the problem. A range of alternative conceptions have been proposed, including “malinformation” and “information disorder.” Recently, some experts have advocated holistic concepts, like “information ecology” or “information and society,” that shift atten-

tion away from individual actors or claims and toward larger social systems. Meanwhile, platforms have developed their own quasi-legalistic argot—such as Meta’s “coordinated inauthentic behavior”—to facilitate governance and enforcement.

There is also a growing set of scholars and commentators who believe the field itself, not just its terminology, must be fundamentally rethought.⁷ Some point out that disinformation and its ilk are elastic notions which tend to reflect the biases of whoever invokes them. Others observe that disinformation isn’t pervasive or influential enough to explain the ills often attributed to it. Several critics have gone so far as to label the disinformation crisis a moral panic, one suffered most acutely by elite groups. On this telling, privileged and expert classes—such as the White liberals who for decades dominated academia and journalism in the United States—have seized upon a perceived surge of disinformation to explain their recent loss of control over the national discourse. This story, rooted in nostalgia for a mythical era of shared truth, offers a comforting, depoliticized morality play: right-thinking in-groups are under siege by ignorant out-groups in the thrall of manipulative (often foreign) bogeymen. The narrative has troubling historical antecedents, such as baseless Cold War-era fears of communist “brainwashing” that led to curtailment of civil liberties in the West.

Despite all these complications and pitfalls, this report begrudgingly embraces the term “disinformation” for three primary reasons. First, it captures a specific, real, and damaging phenomenon: malicious falsehoods are undermining democratic stability and governance around the world. However difficult it may be to identify or define disinformation at the edges, a set of core cases clearly exists and deserves serious attention from policymakers. A paradigmatic example is the “Stop the Steal” movement in the United States. The claim that the 2020 presidential election was stolen is provably false, was put forward with demonstrated bad faith, and has deeply destabilized the country. Second, other phrases have their own problems, and no single term has yet emerged as a clearly better alternative. Third, “disinformation” remains among the most familiar terms for policymakers and other stakeholders who constitute the key audience for this report.

EVALUATION CHALLENGES

Beyond the conceptual issues, policymakers should also be aware of several foundational challenges in assessing the efficacy of disinformation countermeasures. Each of these challenges emerged time and again in the development of this report’s case studies.

- **The underlying problem is hard to measure.** It is hard to know how well a countermeasure works if analysts don’t also know how much impact disinformation has, both before and after the countermeasure is implemented. In fact, countermeasures are only necessary insofar as disinformation is influential to begin with. Unfortunately, experts broadly agree that disinformation (like other forms of

influence) is poorly understood and hard to quantify. A 2021 Princeton University meta-analysis commissioned by Carnegie found that “[e]mpirical research on how influence operations can affect people and societies—for example, by altering beliefs, changing voting behavior, or inspiring political violence—is limited and scattered.”⁸ It specifically noted that “empirical research does not yet adequately answer many of the most pressing questions facing policymakers” regarding the effectiveness of various influence tactics, the role of the medium used (such as specific online platforms), the duration of influence effects, and country-level differences. Until more is known about disinformation itself, the ability to assess countermeasures will remain limited.

- **Success can be defined in multiple ways.** What makes an intervention successful in countering disinformation? An effective intervention might be one that stops someone from embracing a false belief, or discourages people from acting based on false claims, or slows the spread of false information, or protects the integrity of democratic decisionmaking, among other possibilities. All of these effects can be measured over varying time horizons. Additionally, effectiveness is tied to an intervention’s cost, scalability, and the willingness of key stakeholders to facilitate implementation. The risk of blowback is another factor: decisionmakers should consider potential second-, third-, and higher-order effects on the information environment. In short, there is no single way to understand success. Policymakers must decide this for themselves.
- **Policies can coincide, synergize, and conflict with each other.** This report offers discrete evaluations of ten countermeasure types. In reality, multiple kinds of interventions should be implemented at the same time. Simultaneous, interconnected efforts are necessary to address the many complex drivers of disinformation. Policymakers and analysts must therefore avoid judging any one policy option as if it could or should provide a comprehensive solution. An ideal assessment would consider how several interventions can work together, including potential synergies, conflicts, and trade-offs. Such holistic analysis would be extremely difficult to do, however, and is beyond the scope of this report.
- **Subpopulations matter and may react differently.** Many studies of disinformation countermeasures focus on their overall efficacy with respect to the general population, or the “average” person. However, atypical people—those at the tails of the statistical distribution—sometimes matter more. People who consume or share the largest amount of disinformation, hold the most extreme or conspiratorial views, have the biggest influence in their social network, or harbor the greatest propensity for violence often have disproportionate impact on society. Yet these

tail groups are harder to study. Policymakers should take care not to assume that interventions which appear generally effective have the same level of impact on important tail groups. Conversely, interventions that look ineffective at a population level may still be able to influence key subpopulations.

- **Findings may not generalize across countries and regions.** The feasibility and impact of an intervention can vary from place to place. For example, the United States is more polarized than most other advanced democracies, and it faces greater constitutional constraints and government gridlock. On the other hand, the United States has outsized influence over the world's leading social media platforms and possesses relatively wealthy philanthropic institutions and, at the national level, a robust independent press. These kinds of distinctive characteristics will shape what works in the United States, while other countries must consider their own national contexts. Unfortunately, much of the available research focuses on the United States and a handful of other wealthy Western democracies. This report incorporates some examples from other countries, but geographic bias remains present.

These evaluation challenges have no easy solutions. Researchers are working to fill knowledge gaps and define clearer policy objectives, but doing so will take years or even decades. Meanwhile, policymakers must somehow forge ahead. Ideally, they will draw upon the best information available while remaining cognizant of the many unknowns. The following case studies are designed with those twin goals in mind.

CASE STUDY 1

SUPPORTING LOCAL JOURNALISM

DESCRIPTION AND USE CASES

Many analysts have called for investing in local journalism—especially print and digital media—as a way to counter disinformation. The hope is that high-quality local journalism can inform democratic deliberation, debunk false claims, and restore the feelings of trust and community that help to keep conspiracy theories at bay.⁹ More specifically, new financial investments would aim to halt or reverse the industry’s long-term financial deterioration. Local newspapers and other outlets have seen steady declines in ad revenue and readership for the last two decades, as the internet gave birth to more sophisticated forms of digital advertising and alternative sources of free information. According to one count, a fourth of the newspapers operating in the United States in 2004 had closed by the end of 2020.¹⁰ The COVID-19 pandemic accelerated this trend, causing widespread layoffs across print, broadcast, radio, and digital outlets.¹¹ Such challenges have not been limited to the United States or Western countries: for example, COVID-19 “ravaged the revenue base” of Nigerian media organizations, according to one local publisher.¹²

New funding for local journalism could come from governments, philanthropists, commercial sources, or a combination of these. One model for government funding is the New Jersey Civic Information Consortium, a state-supported nonprofit. The consortium receives money from government and private sources, then disburses grants to projects that promote the “quantity and quality of civic information.”¹³ The use of a nonprofit intermediary aims to reduce the risk that government officials would leverage state funds to influence news coverage.¹⁴ Another model is for governments to use tax exemptions and other policy tools

to financially boost the journalism industry without directly subsidizing it.¹⁵ In the United Kingdom, newspapers, books, and some news sites are exempt from the Value-Added Tax because of their public benefit.¹⁶ In Canada, people who purchase a digital news subscription can claim a tax exemption.¹⁷ Australia has taken another approach by passing legislation that empowers news publishers to jointly negotiate for compensation when platforms like Facebook and Google link to their content.¹⁸ Other advocates have proposed a tax on digital advertising that would be used to support journalism.¹⁹

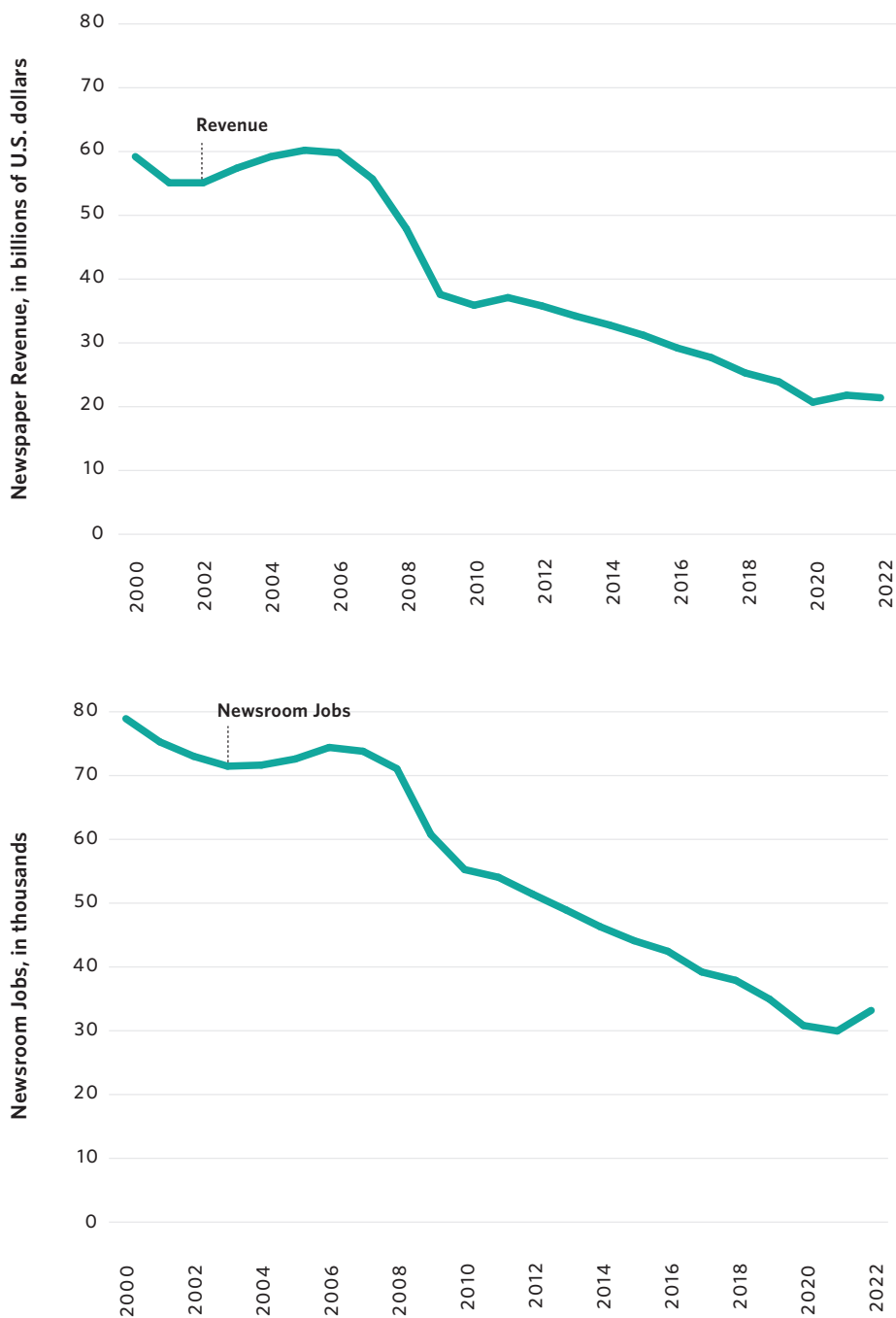
KEY TAKEAWAYS:

There is strong evidence that the decline of local news outlets, particularly newspapers, has eroded civic engagement, knowledge, and trust—helping disinformation to proliferate. Bolstering local journalism could plausibly help to arrest or reverse such trends, but this has not been directly tested. Cost is a major challenge, given the expense of quality journalism and the depth of the industry’s financial decline. Philanthropy can provide targeted support, such as seed money for experimentation. But a long-term solution would probably require government intervention and/or alternate business models. This could include direct subsidies (channeled through nongovernmental intermediaries) or indirect measures, such as tax exemptions and bargaining rights.

KEY SOURCES:

- “INN Index 2022: Enduring in Crisis, Surging in Local Communities,” Institute for Nonprofit News, July 27, 2022, <https://inn.org/research/inn-index/inn-index-2022/>.
- Penelope Muse Abernathy, “News Deserts and Ghost Newspapers: Will Local News Survive?,” University of North Carolina, 2020, <https://www.usnewsdeserts.com/reports/news-deserts-and-ghost-newspapers-will-local-news-survive/>.
- Emily Bell, “Facebook Is Eating the World: It’s the End of the News as We Know It,” *Columbia Journalism Review*, March 7, 2016, <https://www.cjr.org/60th/facebook-is-eating-the-world-emily-bell-end-of-news-as-we-know-it.php>.

Figure 1. Decline of U.S. Newspapers Since 2000



Sources: "Newspapers Fact Sheet," Pew Research Center, November 15, 2023, <https://www.pewresearch.org/journalism/fact-sheet/newspapers/>; Mason Walker, "U.S. Newsroom Employment Has Fallen 26% Since 2008," Pew Research Center, July 13, 2021, <https://www.pewresearch.org/short-reads/2021/07/13/u-s-newsroom-employment-has-fallen-26-since-2008/>; "Occupational Employment and Wage Statistics," U.S. Bureau of Labor Statistics, April 27, 2023, <https://www.bls.gov/oes/tables.htm>.

Philanthropic support for local journalism can also come in various forms. Not-for-profit news outlets in North America currently get about half of their revenue from foundation grants, but national and global outlets receive more than two-thirds of these grant dollars.²⁰ To bolster local outlets, a greater portion of grants could be redirected to them. The next largest source of funding for nonprofit newsrooms is individual gifts, which make up about 30 percent of revenue and primarily come from donations of \$5,000 or more.²¹ However, small-dollar donations are growing; NewsMatch, a U.S. fundraising effort, encourages audiences to donate to local media organizations and matches individual donations with other sources of philanthropy. NewsMatch has raised more than \$271 million since 2017.²²

Multiple government, philanthropic, or commercial revenue streams can be combined in novel ways, as illustrated by Report for America. The initiative raised \$8 million in 2022 to place reporting fellows in local newsrooms.²³ A relatively small portion, about \$650,000, was taxpayer money from the Corporation for Public Broadcasting.²⁴ The remainder came from foundations and technology companies, matched dollar-for-dollar by contributions split between the local newsrooms themselves and other local funders.

HOW MUCH DO WE KNOW?

Research is clear that the decline of local journalism is associated with the drivers of disinformation. However, the inverse proposition—that greater funding for local journalists will reduce disinformation—does not automatically follow and has not been empirically tested.

Going forward, decisionmakers and scholars could study the link between disinformation and the health of local media outlets more closely by monitoring and evaluating the impact of local news startups on a variety of metrics related to disinformation, such as polarization, professed trust in institutions like the media and government, civic engagement and voter turnout, and susceptibility to online rumors.

HOW EFFECTIVE DOES IT SEEM?

Studies suggest at least two mechanisms whereby the decline of local media outlets can fuel the spread of disinformation.

First, the decline contributes to civic ignorance and apathy as voters become less informed about the issues, candidates, and stakes in local elections. Research indicates that reduced access to local news is linked to lower voter turnout and civic engagement as well as increased corruption and mismanagement. At least one longitudinal study also finds a relationship between the decline of local news, on the one hand, and diminished civic awareness and en-

gagement on the other hand.²⁵ These conditions ultimately erode public trust, which can increase belief in misinformation and conspiracy theories.²⁶ Conversely, scholarship has shown that strong media is linked to robust civic participation. Many studies correlate the existence of local newspapers

with higher turnout in local elections. And, at an individual level, a person's consumption of local political news is associated with higher likelihood to vote.²⁷ These patterns can be seen in a variety of electoral contexts—including downballot and judicial elections—and across historical periods, despite changing technology.²⁸ A study of U.S. history from 1869 to 2004 found that a community's civic participation rose when its first newspaper was created, and that this connection persisted even after the introduction of radio and television.²⁹

Second, when local media disappears, lower-quality information sources can fill the gap as people look elsewhere for information. Social media has emerged as a primary alternative.³⁰ Although social media platforms contain plenty of accurate and authoritative voices, they also create opportunities for low-quality and hyperpartisan personalities and outlets (some of which pose as local newspapers) that spread misleading, divisive content.³¹ Indeed, research shows a connection between the decline of local media and the rise of polarization. For example, one study found that communities that lost their local newspaper became more polarized as voters replaced information from local media with more partisan cues picked up elsewhere, such as national cable TV.³² To be sure, polarizing content should not be equated with disinformation. Nevertheless, most analysts believe the two are linked: as voters drift away from the “mainstream” of the political spectrum—often, but not always, toward the right—they may become more accepting of less credible alternative media sources and misleading claims that align with their partisan preferences and demonize political opponents.³³

Given the evidence that local media declines breed susceptibility to disinformation, it is reasonable to predict that efforts to bolster local media could have the opposite effect. However, that prediction has not yet been empirically tested. It is possible, for example, that people who have drifted from traditional local journalism toward social media as an information source might have developed new habits that would be difficult to reverse. Likewise, communities that have suffered a general loss of civic engagement and trust due to the decline of local media might now have less interest or faith in a startup newsroom than they previously would have.

***When local media disappears,
lower-quality information sources
can fill the gap as people look
elsewhere for information.***

HOW EASILY DOES IT SCALE?

Reversing the decline of local journalism is an extremely costly proposition, at least in the United States, because the scale of downsizing has been so large. A Georgetown University study found that newspapers employed 150,000 fewer people in 2022 compared to the 1980s—a decline of 63 percent. Although web publishers have replaced about half of those jobs, replacing the rest would require tremendous investment. For example, the American Journalism Project raised over \$100 million to partially fund thirty-three nonprofit newsrooms—a small fraction of the 2,100 newsrooms that closed in the United States in the past two decades.³⁴ *Washington Post* columnist Perry Bacon Jr. estimated in 2022 that it would cost at least \$10 billion per year to hire 87,000 new journalists—that is, to ensure that each U.S. congressional district had 200 journalists, plus operational support.³⁵ More localized coverage could be even costlier. In 2022, Democracy Fund created a calculator to estimate the total cost of meeting the information needs of every community in the United States. Hiring several reporters to cover crucial issues in each township and municipality would cost \$52 billion per year.³⁶

Philanthropy can provide targeted investments in particularly needy areas—for example, communities too small or poor to sustain local media on their own—and offer seed money to run experiments. But given the sums required, a large-scale solution would demand some combination of long-term government support, new journalistic business models, or other structural changes in the marketplace. The Australian bargaining law provides one promising case study. While critics said the approach would be unlikely to generate much revenue and would mostly benefit large publishers, an Australian government review found that Google and Meta reached thirty agreements with publications of varying size, including some groups of outlets. In its first year, the law raised more than \$140 million for these outlets, much of which was used to hire new journalists and purchase equipment.³⁷ Similar schemes are now being implemented in Canada and under consideration in California—though these efforts, like the Australia law, have faced strong initial pushback from big tech companies.³⁸

CASE STUDY 2

MEDIA LITERACY EDUCATION

DESCRIPTION AND USE CASES

Increasing individuals' media literacy through education and training is one of the most frequently recommended countermeasures against disinformation.³⁹ Proponents argue that “media literacy and critical thinking are the first barrier to deception” and that teaching people these skills therefore enables them to better identify false claims.⁴⁰ The National Association for Media Literacy Education defines media literacy as “the ability to access, analyze, evaluate, create, and act using all forms of communication.” However, scholars point to conceptual confusion around the term, and practitioners take many different approaches.⁴¹ Common goals include instilling knowledge of the media industry and journalistic practices, awareness of media manipulation and disinformation techniques, and familiarity with the internet and digital technologies.

Media literacy education initiatives target a range of different audiences, occur in multiple settings, and use a variety of methods—including intensive classroom-based coursework as well as short online videos and games. Many programs focus on children and adolescents,⁴² with research suggesting that young people are less familiar with the workings of the internet and digital media and more susceptible to online hoaxes and propaganda than commonly assumed.⁴³ For example, a 2016 study of over 7,800 students found many failed to distinguish sponsored content and untrustworthy websites in search results.⁴⁴ Public education is therefore one major vehicle to reach large numbers of people early in their lives, alongside other kinds of youth programs. Aspects of media literacy have long been

embedded in general education and liberal arts curricula in advanced democracies, especially in subjects that emphasize critical reading and thinking, such as language arts, essay writing, civics, and rhetoric. Public libraries have also historically promoted media literacy.

KEY TAKEAWAYS:

There is significant evidence that media literacy training can help people identify false stories and unreliable news sources. However, variation in pedagogical approaches means the effectiveness of one program does not necessarily imply the effectiveness of another. The most successful variants empower motivated individuals to take control of their media consumption and seek out high-quality information—instilling confidence and a sense of responsibility alongside skills development. While media literacy training shows promise, it suffers challenges in speed, scale, and targeting. Reaching large numbers of people, including those most susceptible to disinformation, is expensive and takes many years.

KEY SOURCES:

- Monica Bulger and Patrick Davison, “The Promises, Challenges, and Futures of Media Literacy,” *Data & Society*, February 21, 2018, <https://datasociety.net/library/the-promises-challenges-and-futures-of-media-literacy>.
- Géraldine Wuyckens, Normand Landry, and Pierre Fastrez, “Untangling Media Literacy, Information Literacy, and Digital Literacy: A Systematic Meta-review of Core Concepts in Media Education,” *Journal of Media Literacy Education* 14 (2022), <https://digitalcommons.uri.edu/cgi/viewcontent.cgi?article=1531&context=jmle>.
- Erin Murrock, Joy Amulya, Mehri Druckman, and Tetiana Liubyva, “Winning the War on State-Sponsored Propaganda: Gains in the Ability to Detect Disinformation a Year and a Half After Completing a Ukrainian News Media Literacy Program,” *Journal of Media Literacy Education* 10 (2018): <https://digitalcommons.uri.edu/cgi/viewcontent.cgi?article=1361&context=jmle>.

Not all media literacy programs target young people. After all, people don't necessarily age out of their susceptibility to disinformation; in fact, older individuals seem more likely to share false stories on Facebook.⁴⁵ Media literacy training for adults may happen at libraries, senior citizen centers, recreational events, or professional settings. Civil society and government agencies have also run public awareness campaigns and released gamified education tools. For example, Sweden established a Psychological Defence Agency in 2022. Its responsibilities include leading "training, exercises and knowledge development" to help residents "identify and counter foreign malign information influence, disinformation and other dissemination of misleading information directed at Sweden."⁴⁶

One valuable case study is the International Research and Exchanges Board (IREX)'s Learn to Discern program, which has used a "train the trainers" approach in Ukraine and a number of other countries since 2015. This program equips volunteers to deliver a media literacy curriculum to members of their community.⁴⁷ Reaching more vulnerable adults (for example, racial and ethnic minorities and those with fewer economic resources, less education, or less experience with the internet) is a policy priority for governments focused on media literacy.⁴⁸

HOW MUCH DO WE KNOW?

The body of scholarship on media literacy is large relative to most other disinformation countermeasures. For example, a 2022 literature review on digital literacy—one component of media literacy—found forty-three English-language studies since 2001, with thirty-three of these published since 2017, when interest in the topic swelled.⁴⁹ The existence of dedicated journals and conferences is another indicator of growth in this subfield. For example, the National Association for Media Literacy Education published the first issue of the *Journal of Media Literacy Education* in 2009.⁵⁰ Other major repositories of research on media literacy include a database maintained by the United Nations Alliance of Civilizations.⁵¹

Review of this literature shows that specific media literacy approaches have a strong theoretical basis and a large body of experimental evidence. However, variation in pedagogical approaches means the effectiveness of one program does not necessarily imply the effectiveness of another.⁵² Moreover, the lack of robust mechanisms for collecting data on classroom activities is a recognized gap. In 2018, the Media Literacy Programme Fund in the United Kingdom (considered a leader in media literacy education) cited grants to support evaluation as a priority.⁵³ Since then, several studies have conducted real-time evaluation and sought to measure lasting improvements in student performance. Additional studies could expand the menu of possible approaches to evaluation; also useful would be to examine further the effectiveness of media literacy training for atypical individuals at the extremes, such as those who are especially motivated by partisanship, conspiracy theories, or radical ideologies.

HOW EFFECTIVE DOES IT SEEM?

There is significant evidence that media literacy training can help people identify false stories and unreliable news sources.⁵⁴ Scholars sometimes refer to this as inoculation, because “preemptively exposing, warning, and familiarising people with the strategies used in the production of fake news helps confer cognitive immunity when exposed to real misinformation.”⁵⁵ One experiment found that playing an online browser game designed to expose players to six different disinformation strategies reduced subjects’ susceptibility to false claims, especially among those users who were initially most vulnerable to being misled. Such laboratory findings are bolstered by studies of larger, real-world interventions. An evaluation of IREX’s Learn to Discern program found durable increases in good media consumption habits, such as checking multiple sources, lasting up to eighteen months after delivery of the training.⁵⁶ Other studies support teaching students to read “laterally”—using additional, trusted sources to corroborate suspect information.⁵⁷

Because media literacy comes in many forms, it is important to assess which variants are most effective at reducing belief in false stories so trainers and educators can prioritize them. Research suggests that the most successful variants empower motivated individuals to take control of their media consumption and seek out high-quality information. This has been described as “actionable skepticism,” or sometimes simply as “information literacy.”⁵⁸ For example, a 2019 review in *American Behavioral Scientist* examined various factors that might enable someone to recognize false news stories. They found that people’s “abilities to navigate and find information online that is verified and reliable”—for example, differentiating between an encyclopedia and a scientific journal—was an important predictor. In contrast, subjects’ understanding of the media industry and journalistic practices or their self-reported ability to “critically consume, question, and analyze information” were not predictive.⁵⁹ Later research based on survey data also supported these findings.⁶⁰

Importantly, multiple studies have shown that effective media literacy depends not only on people’s skills but also on their feelings and self-perceptions. Specifically, individuals who feel confident in their ability to find high-quality news sources, and who feel responsible for proactively doing so, are less likely to believe misleading claims. This factor is often called

The most successful variants empower motivated individuals to take control of their media consumption and seek out high-quality information.

an individual’s “locus of control,” and it has been identified as important in studies of multiple nationally and demographically diverse populations.⁶¹ People who purposefully curate their information diet are less likely to be misled; passive consumers, on the other hand, are more vulnerable. However, this may be truer of typical news

consumers than of outliers like extremists and very motivated partisans. The latter groups might self-report confidence in curating their media diet while nevertheless selecting for misleading, radical, or hyper-partisan sources.

A growing body of recent literature based on large-scale classroom studies shows how specific techniques can provide news consumers with greater agency and ability to seek out accurate information.⁶² Whereas past forms of online media literacy education often focused on identifying markers of suspicious websites—like typographical errors or other indicators of low quality—these signs are less useful in the modern information environment, where sources of misinformation can have the appearance of high production value for low cost.⁶³ Recent studies have shown that lateral reading is more effective.⁶⁴ In one study of students at a public college in the northeastern United States, only 12 percent of subjects used lateral reading before receiving training on how to do so; afterward, more than half did, and students showed an overall greater ability to discern true claims from fictional ones.⁶⁵ A similar study on university students in California found these effects endured after five weeks.⁶⁶ Another one-day exercise with American middle school students found that students had a difficult time overcoming impressions formed from “superficial features” on websites and should be trained to recognize different types of information sources, question the motivation behind them, and—crucially—compare those sources with known trustworthy sites.⁶⁷

Teaching people to recognize unreliable news sources and common media manipulation tactics becomes even more effective when participants are also able to improve their locus of control, according to academic research and program evaluations. In a study of media literacy among 500 teenagers, researchers found that students with higher locus of control were more resilient against false stories. In another study based on survey data, researchers found that individuals who exhibited high locus of control and the ability to identify false stories were more likely to take corrective action on social media, such as reporting to the platform or educating the poster.⁶⁸ (The participatory nature of social media increases the importance of educating users not only on how to recognize untrustworthy content but also on how to respond to and avoid sharing it.⁶⁹)

Evaluations of IREX’s Learn to Discern program in Ukraine and a similar program run by PEN America in the United States shed further light on locus of control. These curricula’s focus on identifying untrustworthy content led subjects to become overly skeptical of *all* media. While trainees’ ability to identify disinformation and their knowledge of the news media increased, their locus of control changed only slightly. Ultimately, trainees’ ability to identify accurate news stories did not improve, and they remained distrustful of the media as a whole.⁷⁰ A major challenge, then, is news consumers who feel under threat from the information environment rather than empowered to inform themselves. One potential intervention point could be social media platforms, which can provide tools and make

other design choices to help users compare on-platform information with credible external sources (see case study 4). This could reinforce users' locus of control while assisting them in exercising it.

Educators should be mindful of media literacy expert Paul Mihailidis's warning that "critical thought can quickly become cynical thought."⁷¹ In a 2018 essay, media scholar danah boyd argued that individuals who are both cynical about institutions and equipped to critique them can become believers in, and advocates for, conspiracy theories and disinformation. To avoid this trap, media literacy education must be designed carefully. This means empowering people to engage with media critically, constructively, and discerningly rather than through the lenses of undifferentiated paranoia and distrust.⁷²

HOW EASILY DOES IT SCALE?

While media literacy training shows promise, it suffers challenges from speed, scale, and targeting. Many approaches will take years to reach large numbers of people, including many vulnerable and hard-to-reach populations. Attempts to reach scale through faster, leaner approaches, like gamified online modules or community-based efforts to train the trainers, are highly voluntary and most likely to impact already motivated individuals rather than large percentages of the public.

Many media literacy projects are not particularly expensive to deliver to small audiences. However, achieving wide impact requires high-scale delivery, such as integrating media literacy into major institutions like public education—a costly proposition. When a proposed 2010 bill in the U.S. Congress, the Healthy Media for Youth Act, called for \$40 million for youth media literacy initiatives, leading scholars deemed the amount insufficient and advocated for larger financial commitments from the government, foundations, and the private sector.⁷³

Once the resources and curricula are in place, it will still take time to develop necessary infrastructure to implement large-scale media literacy programs. For example, hiring skilled educators is a critical yet difficult task. Studies from the European Union (EU) and South Africa both identified major deficiencies in teachers' own abilities to define core media literacy concepts or practice those concepts themselves.⁷⁴

CASE STUDY 3

FACT-CHECKING

DESCRIPTION AND USE CASES

Fact-checking, in this report, refers broadly to the issuance of corrective information to debunk a false or misleading claim. A 2020 global survey by Carnegie identified 176 initiatives focused on fact-checking and journalism, while the Duke University Reporters' Lab counted more than 400 active fact-checking efforts across more than 100 countries in 2023.⁷⁵ These initiatives come in many different forms. They include dedicated, stand-alone organizations, such as Snopes, as well as fact-checkers integrated into newspapers and TV programs. Some prioritize political claims, like the *Washington Post's* "Fact Checker" and the website PolitiFact. Others address health claims, like the CoronaVirusFacts/DatosCoronaVirus Alliance Database led by the International Fact-Checking Network at the Poynter Institute.⁷⁶

Collaborative fact-checking models uniting the efforts of several organizations have also emerged, like Verificado 2018, an effort to collect rumors and disinformation circulating on WhatsApp during the 2018 Mexican elections and deliver corrections through private messaging.⁷⁷ Projects like this attempt to quickly reach a large audience through a medium people already use. Other initiatives in multiple countries have attempted to crowdsource from citizen fact-checkers.

KEY TAKEAWAYS:

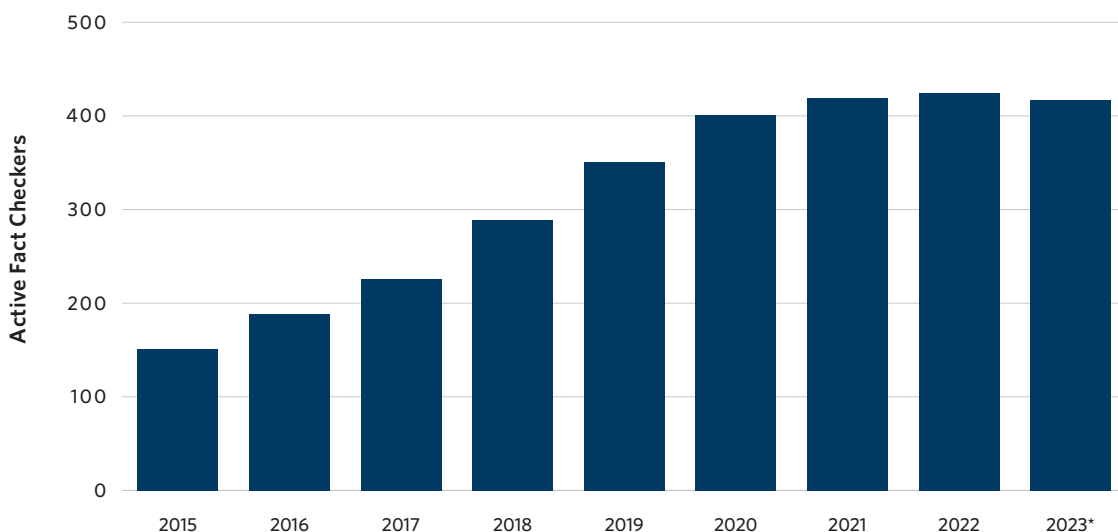
A large body of research indicates that fact-checking can be an effective way to correct false beliefs about specific claims, especially for audiences that are not heavily invested in the partisan elements of the claims. However, influencing factual beliefs does not necessarily result in attitudinal or behavioral changes, such as reduced support for a deceitful politician or a baseless policy proposal. Moreover, the efficacy of fact-checking depends a great deal on contextual factors—such as wording, presentation, and source—that are not well understood. Even so, fact-checking seems unlikely to cause a backfire effect that leads people to double down on false beliefs. Fact-checkers face a structural disadvantage in that false claims can be created more cheaply and disseminated more quickly than corrective information; conceivably, technological innovations could help shift this balance.

KEY SOURCES:

- Brendan Nyhan, Ethan Porter, Jason Reifler, Thomas Wood, “Taking Fact-Checks Literally But Not Seriously? The Effects of Journalistic Fact-Checking on Factual Beliefs and Candidate Favorability,” *Political Behavior* 42 (2019): <https://link.springer.com/article/10.1007/s11109-019-09528-x>.
- Thomas Wood and Ethan Porter, “The Elusive Backfire Effect: Mass Attitudes’ Steadfast Factual Adherence,” *Political Behavior* 41 (2019): <https://link.springer.com/article/10.1007/s11109-018-9443-y>.
- Emily Thorson, “Belief Echoes: The Persistent Effects of Corrected Misinformation,” *Political Communication* 33 (2015): <https://www.tandfonline.com/doi/full/10.1080/10584609.2015.1102187>.

In recent years, some social media companies have highlighted fact-checks on their platforms and used the assessments of fact-checkers to inform other policy actions. For example, Meta’s third-party fact-checking program routes Facebook and Instagram posts that contain potential falsehoods to fact-checkers certified through the International Fact-Checking Network and applies a label if the posts are false or disputed.⁷⁸ (For more on social media labeling, see case study 4.) Beyond social media, fact-checks can also be disseminated on dedicated websites or during televised political debates, among other possibilities.

Figure 2. Number of Fact-Checking Outlets Globally, 2015–2023



Source: Mark Stencel, Erica Ryan, and Joel Luther, “Misinformation Spreads, but Fact-Checking Has Leveled Off,” Duke Reporters’ Lab, June 21, 2023, <https://reporterslab.org/misinformation-spreads-but-fact-checking-has-leveled-off>.

Note: *2023 data is as of June 2023.

HOW MUCH DO WE KNOW?

Fact-checking is well-studied—markedly more so than other interventions. Nearly 200 articles related to fact-checking published since 2013 were reviewed for this case study. However, the strong empirical research base also reveals that fact-checking’s effectiveness depends on a complex interplay of multiple factors which remain poorly understood. Research has only begun to probe the specific parameters that apparently affect fact-checking’s impact, such as format, language, and source. Additionally, much of the academic literature on fact-checking comes from laboratory studies based on unrepresentative samples of university students, or from online quizzes based on crowdsourcing platforms like Amazon’s Mechanical Turk—raising questions about the findings’ generalizability. Among other problems, the subjects of such studies may be more interested or engaged with fact-checking content presented to them by experimenters, as compared with members of the general public who encounter such content organically. More research evaluating the longitudinal impact of ongoing fact-checking efforts in a diverse set of real-time, real-world environments is still needed.

HOW EFFECTIVE DOES IT SEEM?

A number of studies suggest that it is easier to cause people to disbelieve false claims but harder to change the behaviors related to those beliefs. For example, international studies have shown fact-checks to have some success at changing beliefs about viral diseases, but they do not always lead to increased intent to receive vaccines or improved public health behaviors.⁷⁹ This disconnect may be especially large for politically charged topics in divided societies. Fact-checking the claims of political figures has limited impact on voters' support for a candidate or policy position—even when the voters can correctly reject false claims.⁸⁰

In general, studies find strong evidence of confirmation bias: subjects are more susceptible to false claims that align with preexisting beliefs or allegiances and are more resistant to fact-checks associated with an opposing political party or its positions.⁸¹ In fact, research suggests that accuracy is not always a top-of-mind issue for news consumers. For example, one 2013 study suggested that individuals put more stock in the perceived trustworthiness (or sincerity) of a corrective source than in the source's actual expertise on the relevant topic.⁸² In another study, right-leaning, U.S.-based participants who were asked to judge the validity of articles tended to provide “expressive” assessments—aimed more at demonstrating their partisan allegiance than at seriously evaluating a source's credibility.⁸³ To be sure, many studies of fact-checking and confirmation bias focus on U.S. audiences, where political polarization is especially strong.⁸⁴ It is possible that partisan barriers to fact-checking are less present in more unified societies.⁸⁵

Some research initially sparked concern that fact-checking might perversely cause audiences to double down on their false beliefs. The term “backfire effect” was initially coined to describe this behavior in a 2010 article by political scientists Brendan Nyhan and Jason Reifler and took root in American public consciousness after the 2016 U.S. presidential election.⁸⁶ However, more recent research (including by Nyhan) suggests that backfiring may be a rare phenomenon.

The efficacy of fact-checks depends on many factors. The precise wording of fact-checks matters, with more straightforward refutations being more effective than nuanced explanations. Additionally, one 2015 study found that a fact-check that provides an alternative “causal explanation for an unexplained event is significantly more effective than a denial even when the denial is backed by unusually strong evidence.”⁸⁷ In other words, replacing a false story with a true story works better than merely refuting the false story. However, many of these factors remain poorly understood; for example, research is inconclusive on whether fact-checks should repeat the false claim being debunked or avoid doing so.

The use of emotion and storytelling in fact-checks is another potentially important but under-researched area. One study found that “narrative correctives,” which embed fact-checks within an engaging story, can be effective—and stories that end on an emotional note, such as fear or anger, work better than those that do not.⁸⁸ Another study suggested that anger and anxiety increase motivated reasoning and partisan reactions, although this did not seem to prevent fact-checks from influencing users.⁸⁹

One of the most important outstanding research areas is the durability of fact-checks: how long is corrective information remembered and believed by the recipient? Studies have reached complicated or conflicting results. Some research, for example, has suggested that a recipient’s increase in knowledge of truthful information may last longer than any change in deeper beliefs or attitudes related to that knowledge.⁹⁰ This finding highlights an important difference between informational knowledge and affective feeling—both of which influence people’s beliefs and behaviors. A 2015 study found evidence that misinformation affected the audience’s sentiment toward public figures even after false claims were immediately debunked.⁹¹

HOW EASILY DOES IT SCALE?

The large number of ongoing fact-checking efforts around the world indicates that this intervention can be undertaken at reasonable expense. Some efforts, such as those incorporated into for-profit journalistic enterprises, may even be self-sustaining—whether on their own or as part of a larger business model. Initiatives like the International Fact-Checking Network have received financial and other support from philanthropists, tech companies, and universities.

Fact-checking does face at least two scaling challenges. First, it often takes much more time and expertise to produce a fact-check than to generate the false content being debunked. So long as fact-checkers face this structural disadvantage, fact-checking cannot be a comprehensive solution to disinformation. Rather than scale up to match the full scope of false claims, fact-checkers must instead do triage. Second, fact-checks require distribution mechanisms capable of competing effectively with the spread of disinformation. This means finding ways to reach the audience segments most vulnerable to disinformation. The faster and the more frequent the fact-checks, the better. Ideally, fact-checking should occur before or at the same time as the false information is presented. But this is no easy task. Given the significant investments already being made to produce fact-checks, funders should ensure that distribution mechanisms are sufficient to fully leverage fact-checkers’ work.

Technological innovation may help to reduce the cost of producing high-quality fact-checks and enable their rapid dissemination. Crowdsourcing methods, such as Twitter's Birdwatch (later renamed Community Notes on X), are one approach that merits further study.⁹² Others have begun to test whether generative AI can be used to perform fact-checks. While today's generative AI tools are too unreliable to produce accurate fact-checks without human supervision, they may nevertheless assist human fact-checkers in certain research and verification tasks, lowering costs and increasing speed.⁹³ Ultimately, both crowdsourcing and AI methods still depend on the availability of authoritative, discoverable facts by which claims can be assessed. Producing this factual baseline—whether through science, journalism, or other knowledge-seeking efforts—is an important part of the fact-checking cycle. This too requires funding.

CASE STUDY 4

LABELING SOCIAL MEDIA CONTENT

DESCRIPTION AND USE CASES

Social media companies are increasingly applying labels to content on their platforms, some of which aim to help users assess whether information is trustworthy. In this report, “labeling” refers to the insertion of relevant context or advisories to inform or influence how content is viewed, though without directly fact-checking it. (For more on fact-checking, see case study 3.)

Labels can be applied to a social media account (for example, identifying it as state-sponsored media or satirical) or to individual posts. When a post links to another source, such as an external website, that source can be labeled (as with so-called nutrition labels that score news outlets by their adherence to journalistic practices). Alternatively, specific content or claims can be labeled—as disputed, potentially outdated, or fast-developing, for instance. Some labels are prominent, use firm language, and require a user to click before seeing or interacting with the content. Other labels are small, discreet, and neutrally worded.

Labels can be positive, like a digital signature that verifies video as authentic or a “verified” badge that purports to confirm an account’s identity. Other labels do not seek to inform users, per se, but rather admonish or “nudge” them to follow good information practices. For example, a user seeking to reshare an article may encounter a message that encourages them to first read the article and/or consider its accuracy; such “friction” in user interfaces seeks to promote more deliberate, reflective behavior. Additionally, many common platform design features can loosely be understood as labels. For example, platforms often display engagement data—such as the number of likes, shares, or views—alongside content. This data can influence users’ perceptions of the content’s accuracy and importance.⁹⁴

Facebook was among the first platforms to label misleading content after public concern about so-called fake news and its influence on the 2016 U.S. presidential election.⁹⁵ Other platforms, including Twitter (now X) and YouTube, have also implemented labels of various kinds—often spurred by major events such as the 2020 U.S. presidential election and the COVID-19 pandemic.

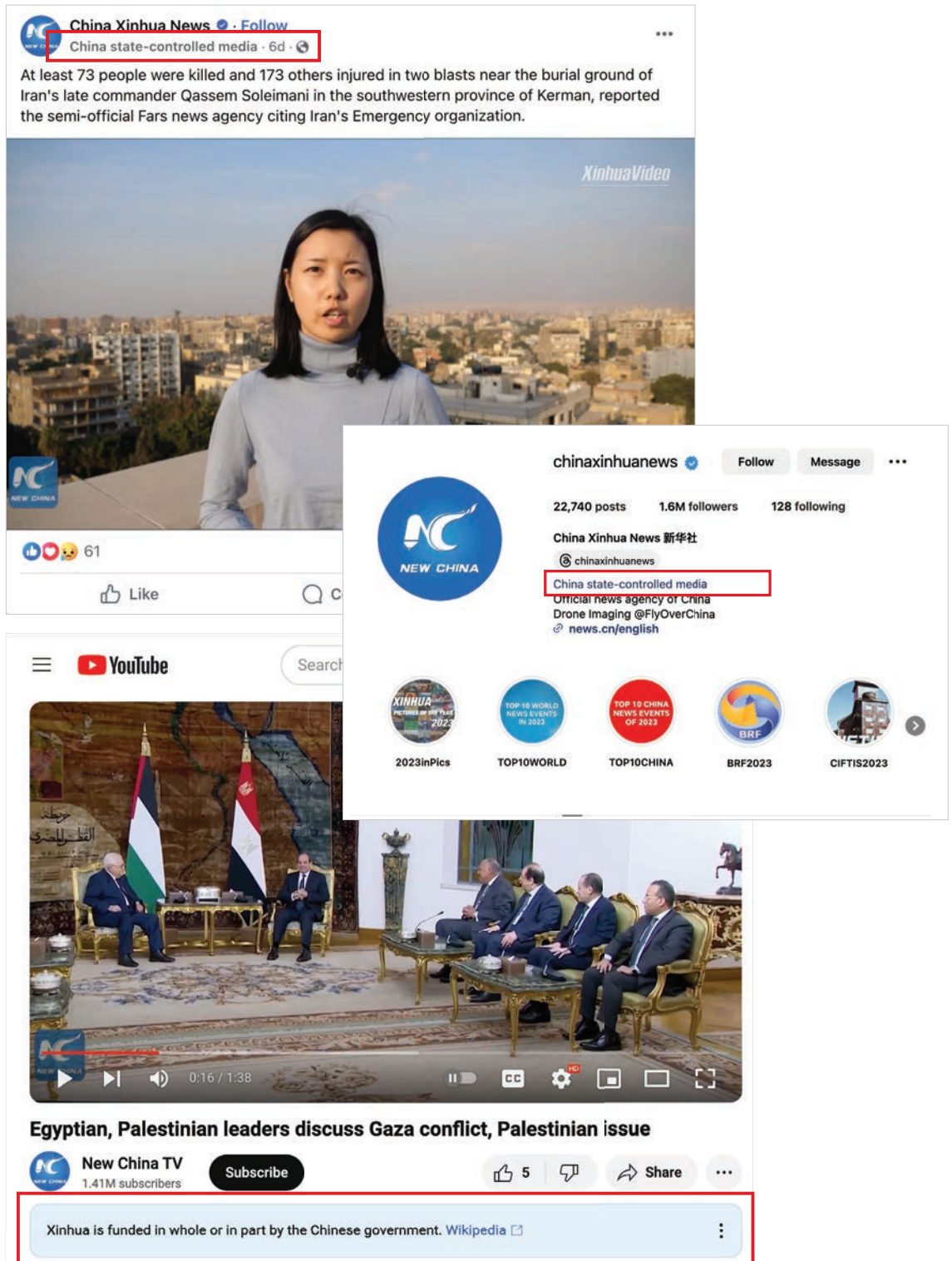
KEY TAKEAWAYS:

There is a good body of evidence that labeling false or untrustworthy content with additional context can make users less likely to believe and share it. Large, assertive, and disruptive labels are the most effective, while cautious and generic labels often do not work. Reminders that nudge users to consider accuracy before resharing show promise, as do efforts to label news outlets with credibility scores. Different audiences may react differently to labels, and there are risks that remain poorly understood: labels can sometimes cause users to become either overly credulous or overly skeptical of unlabeled content, for example. Major social media platforms have embraced labels to a large degree, but further scale-up may require better information-sharing or new technologies that combine human judgment with algorithmic efficiency.

KEY SOURCES:

- Gordon Pennycook, Adam Bear, Evan T. Collins, and David G. Rand, "The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings," *Management Science* 66, no. 11 (November 2020): <https://pubsonline.informs.org/doi/10.1287/mnsc.2019.3478>.
- Kevin Aslett et al., "News Credibility Labels Have Limited Average Effects on News Diet Quality and Fail to Reduce Misperceptions," *Science Advances* 8, no. 18 (May 2022): <https://www.science.org/doi/10.1126/sciadv.abl3844>.
- Gordon Pennycook et al., "Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention," *Psychological Science* 31, no. 7 (2020): <https://journals.sagepub.com/doi/full/10.1177/0956797620939054>.

Figure 3. Screenshots of Facebook, Instagram, and YouTube Labels for State Media



HOW MUCH DO WE KNOW?

The academic literature on labeling is smaller than that on fact-checking but still large compared to other interventions. Social media companies began employing labels in earnest only in 2019, according to a Carnegie database.⁹⁶ Independent studies of social media labels face methodological challenges due to researchers' lack of access to private platform data on how users react to labels, though internal company research occasionally reaches the public domain through leaks, government inquiries, investigative journalism, or voluntary (if selective) disclosure. Laboratory experiments can be helpful, but they do not fully simulate key aspects of real-life social media usage.

HOW EFFECTIVE DOES IT SEEM?

Evidence suggests that large, prominent, and strongly worded labels can sometimes inhibit belief in and spread of false claims. However, other labels appear less effective. For example, studies show that labels which visually stand apart from the adjoining content are more effective than those that blend in. Similarly, labels that deliver a clear warning—for example, by pointing out that the content has previously appeared on an unreliable rumor site—are more effective than those that merely note a claim is “disputed.”⁹⁷

Some internal research by platforms has also indicated that neutrally worded labels may be ineffective and can even lead users to gradually tune them out. During the COVID-19 pandemic, Facebook relied on independent fact-checkers to determine whether COVID-19 content was false or misleading; debunked content would then be labeled as such and algorithmically demoted. But “fact-checkers were unable to review an overwhelming majority of the content in their queue” because of resource limitations, so Facebook also applied neutral labels en masse to all other COVID-19 content. These labels provided context—“COVID-19 vaccines go through many tests for safety and effectiveness and are then monitored closely”—along with a link to authoritative information. According to Meta’s Oversight Board, however, “initial research showed that these labels may have [had] no effect on user knowledge and vaccine attitudes” and “no detectable effect on users’ likelihood

to read, create or re-share” false claims.⁹⁸

Facebook reduced and ultimately rolled back these labels after finding that users became less likely to click through to the information page after repeated label exposure.

Source ratings, either by fact-checkers or other users, have been shown to be effective at reducing engagement on articles

Evidence suggests that large, prominent, and strongly worded labels can sometimes inhibit belief in and spread of false claims. However, other labels appear less effective.

with low scores. Specifically, labels that score a news source's credibility can influence users' willingness to like, comment on, or share posts containing links to news articles. This is a promising finding for projects like NewsGuard, which ranks news sites on a 100-point rubric based on best practices for credible and transparent journalism.⁹⁹ However, empirical studies of NewsGuard have had mixed results. A 2022 study found, on the one hand, that exposure to labels did "not measurably improve news diet quality or reduce misperceptions, on average, among the general population." On the other hand, there was also "suggestive evidence of a substantively meaningful increase in news diet quality among the heaviest consumers of misinformation."¹⁰⁰ This split finding may be considered successful or unsuccessful depending on the specific problem such labels are intended to address.

Recent research suggests that labels containing accuracy nudges, which simply encourage users to consider accuracy before sharing content, are particularly promising. This is perhaps surprising, as one might assume that social media users already seek to consume and share what they deem to be accurate information. Yet studies have highlighted a range of other motives—such as amusement and partisan signaling—that often influence user behavior.¹⁰¹ Despite these psychological tendencies, research suggests that most users nevertheless value accuracy and that labels reminding them to consider accuracy make them less likely to share misinformation.¹⁰² In fact, such labels can reduce—though not eliminate—subjects' inclination to believe and share false stories that align with their political beliefs.¹⁰³

Regardless of how labels are designed and implemented, the nature of the content or speaker being labeled can also influence user response. For example, New York University's Center for Social Media and Politics found that during the 2020 election, tweets by then U.S. president Donald Trump which were labeled as disputed spread *further* than those without a label.¹⁰⁴ This was not true for other politicians' accounts in the sample, suggesting that labels on posts by extremely prominent individuals may perform differently from other labels. Additional research on this topic—for example, exploring figures other than Trump and metrics beyond spread of the post—would be valuable, because extremely prominent individuals are often responsible for a disproportionate amount of disinformation.

Like other interventions, labeling can sometimes have perverse effects. Several studies found evidence that labeling some articles as false or misleading led users to become more credulous toward the remaining unlabeled headlines.¹⁰⁵ Researchers call this the "implied truth effect," because users who become accustomed to seeing labels on some content may mistakenly assume that other content has also been vetted. Such a perverse effect, if prevalent, could have significant consequences: labeling efforts often have limited scope and therefore leave the vast majority of content unlabeled.

Paradoxically, there is also some evidence of an opposite dynamic: fact-checks or warning labels can sometimes increase overall audience skepticism, including distrust of articles that did not receive any rating.¹⁰⁶ This might be called an "implied falsity effect." Little is known

about either effect and why one, or both, of them may be present under varying circumstances. It is possible that geographical, topical, or other unidentified factors may influence the effectiveness of labels and the risk of unintended consequences.¹⁰⁷ Moreover, different audiences can respond differently to the same label.

Finally, it is worth remembering that labels explicitly focused on truth or reliability are not the only ways that platform interfaces actively shape how users perceive social media content. One study found that labeling posts with engagement metrics—such as number of “likes”—makes people more likely to share low-credibility, high-engagement posts.¹⁰⁸ Researchers should continue to explore the influence of general user interface design on disinformation, including whether and how common design elements may alter the efficacy of interventions like labeling.

HOW EASILY DOES IT SCALE?

Major platforms’ embrace of labels has shown that they can be scaled to a significant degree. Labeling, in its various forms, has emerged as the dominant way that social media companies adjust their platforms’ design and functionality to counter disinformation and other kinds of influence operations. A Carnegie database of interventions announced by major platforms between 2014 and 2021 found a surge in labeling and redirection (a related measure) since 2019, with 77 of 104 total platform interventions falling into these two categories.¹⁰⁹ Labels offer platforms a way of addressing disinformation without flatly banning or demoting content, actions that impinge more on users’ freedoms and tend to inspire stronger backlash. As a result of platforms’ experimentation with labels, technical barriers—such as load latency, user friction, and so forth—have been addressed or managed.

However, labeling still carries some of the scaling limitations of fact-checks. Meta’s experience with labeling COVID-19 information illustrates one of the choices facing platforms. They can rely on humans to apply more specific, opinionated, and ultimately effective labels to a smaller amount of content, or they can have algorithms automatically label more content with comparatively cautious, generic labels that tend to be less effective and are sometimes counterproductive. Technological innovations could help to further combine both techniques, as better algorithms do more labeling work under human supervision and/or empower humans to label more efficiently. Such innovations would test platforms’ willingness to apply strong labels to large amounts of content, potentially angering users who disagree with the labels. Future studies can continue to examine the specifics of labels and probe the platforms’ processes for applying them.¹¹⁰

The increasing number of platforms presents another scaling challenge. Content that is labeled on one platform may not be labeled on another. While some platforms shun labels based on an overall strategy of minimal content moderation, other platforms lack sufficient resources or simply haven't faced the same public pressure as larger companies to confront disinformation. Outside organizations could explore whether prodding smaller platforms and offering them resources—such as technology, data, and best practices—might encourage more labeling.

CASE STUDY 5

COUNTER-MESSAGING STRATEGIES

DESCRIPTION AND USE CASES

Counter-messaging, in this report, refers to truthful communications campaigns designed to compete with disinformation at a narrative and psychological level instead of relying solely on the presentation of facts. Counter-messaging is premised on the notion that evidence and logic aren't the only, or even the primary, bases of what people believe. Rather, research has shown that people more readily accept claims which jibe with their preexisting worldviews and accepted stories about how the world works, especially if framed in moral or emotional terms.¹¹¹ Moreover, claims are more persuasive when the messenger is a trusted in-group member who appears to respect the audience members and have their best interests at heart. While such factors often facilitate the spread of disinformation, counter-messaging campaigns seek to leverage them in service of truthful ideas.

In a sense, counter-messaging is no different from ordinary political communication, which routinely uses narratives, emotion, and surrogate messengers to persuade. But counter-messaging is sometimes implemented with the specific goal of countering disinformation—often because purely rational appeals, like fact-checking, seem not to reach or have much impact on hard-core believers of false claims. By changing the narrative frame around an issue and speaking in ways designed to resonate, counter-messaging aims to make audiences more open to facts and less ready to accept sensational falsehoods.

KEY TAKEAWAYS:

There is strong evidence that truthful communications campaigns designed to engage people on a narrative and psychological level are more effective than facts alone. By targeting the deeper feelings and ideas that make false claims appealing, counter-messaging strategies have the potential to impact harder-to-reach audiences. Yet success depends on the complex interplay of many inscrutable factors. The best campaigns use careful audience analysis to select the most resonant messengers, mediums, themes, and styles—but this is a costly process whose success is hard to measure. Promising techniques include communicating respect and empathy, appealing to prosocial values, and giving the audience a sense of agency.

KEY SOURCES:

- Jacob Davey, Henry Tuck, and Amarnath Amarasingam, “An Imprecise Science: Assessing Interventions for the Prevention, Disengagement and De-radicalisation of Left and Right-Wing Extremists,” Institute for Strategic Dialogue, 2019, <https://www.isdglobal.org/isd-publications/an-imprecise-science-assessing-interventions-for-the-prevention-disengagement-and-de-radicalisation-of-left-and-right-wing-extremists/>.
- Rachel Brown and Laura Livingston, “Counteracting Hate and Dangerous Speech Online: Strategies and Considerations,” Toda Peace Institute, March 2019, https://toda.org/assets/files/resources/policy-briefs/t-pb-34_brown-and-livingston_counteracting-hate-and-dangerous-speech-online.pdf.
- Benjamin J. Lee, “Informal Countermessaging: The Potential and Perils of Informal Online Countermessaging,” *Studies in Conflict & Terrorism* 42 (2018): <https://www.tandfonline.com/doi/full/10.1080/1057610X.2018.1513697>.

One example comes from Poland, where xenophobia toward migrants from the Middle East during the Syrian civil war was fueled in part by false stories of disease and criminality.¹¹² A Polish counter-messaging campaign called Our Daily Bread featured a video of refugees and other marginalized people baking bread, a cherished Polish activity. Rather than presenting facts and evidence about the impact of migration on Polish society or refuting

false stories about migrants, the video instead used personal vignettes, evocative imagery, and unifying words. The video attracted significant media attention and was viewed more than 1 million times in the first day after its release.¹¹³ Similarly, many efforts to promote COVID-19 vaccines and counter disinformation about them employed themes of personal responsibility. Other such efforts focused on recruiting local doctors as messengers, based on the premise that many people trust their family doctors more than national authorities.¹¹⁴ Vaccine-related public messaging campaigns also partnered with Christian, Jewish, and Muslim faith leaders to reach religious communities in Israel, the United Kingdom, and the United States.¹¹⁵

As these examples indicate, counter-messaging is not always exclusively aimed at countering false claims; other common objectives include promoting desirable behaviors, bolstering social cohesion, and rallying support for government policies. Many initiatives have sought specifically to thwart terrorist recruitment under the banner of “countering violent extremism” and “deradicalization.” For example, the Redirect Method developed by Jigsaw and Moonshot used digital advertising to steer individuals searching for extremist content toward “constructive alternate messages.”¹¹⁶ Other approaches have used one-on-one online conversations or in-person mentorship relationships to dissuade those showing interest in extremism.¹¹⁷ While many of these efforts were designed to address Islamic extremists, they have also been applied to White supremacist and other hate groups.

HOW MUCH DO WE KNOW?

For decades, disciplines such as social psychology, political science, communications, advertising, and media studies have researched issues relevant to counter-messaging. Fields that have themselves been subject to persistent disinformation—such as public health and climate science—have also devoted a great deal of attention to counter-messaging in recent years. Efforts to study and suppress hate and extremist groups are particularly relevant, because such groups often employ disinformation.¹¹⁸ Nevertheless, these bodies of knowledge, though replete with useful insights, have generally not used disinformation as their primary frame for evaluating the efficacy of counter-messaging. This leaves us to rely on analogies and parallels rather than direct evidence.

The relevant literature highlights how hard it is to assess the impact of any form of persuasion. For example, many studies of COVID-19-related counter-messages measured changes in subjects’ reported attitudes or beliefs but were unable to verify whether those shifts persisted or led to behavioral changes.¹¹⁹ Studies based on surveys or laboratory experiments are common, but these do not fully capture how audiences react in more natural settings. In the field of countering violent extremism, practitioners report lacking the expertise or resources to evaluate the impact of their work beyond using social media engagement

metrics and their gut instinct.¹²⁰ A review of online counter-extremism interventions similarly found “virtually all” of the evaluations included in the study measured processes, like social media engagement, not outcomes. The review offered several proposals for more impact-based assessments, such as the inclusion of calls to action like contacting a hotline, which can be quantified as a sign of behavior.¹²¹

HOW EFFECTIVE DOES IT SEEM?

The core insight of counter-messaging—that communications tailored to the narrative and psychological needs of a specific audience are more effective than generic, purely fact-based approaches—is well-established.¹²² Beyond this basic premise, however, it is difficult to generalize about counter-messaging because of the intervention’s breadth, diversity, and overlap with ordinary politics. Some forms seem capable of affecting individuals’ beliefs and, more rarely, influencing the behaviors informed by those beliefs. Yet success may often depend on

the interplay of a large number of factors that can be difficult to discern or control. A granular understanding of the audience should, in theory, enable the selection of mediums, messengers, messages, styles, and tones most likely to resonate with them.¹²³ In practice, developing this audience understanding is a difficult task and determining the best communication approaches is an evolving science at best.

One theme that emerges from many assessments of counter-messaging . . . is the importance of communicating respect and empathy.

One theme that emerges from many assessments of counter-messaging, including public health and counter-extremism interventions, is the importance of communicating respect and empathy. People are often put off by the sense that they are being debated or chastised.¹²⁴ For example, counselors working with White supremacists had the most success in changing subjects’ views through sustained dialogue that avoided moral judgement.¹²⁵ Encouraging empathy toward others, such as religious minorities or immigrants, can also be effective; one study found that such messages make individuals more likely to delete their previous hate speech and less likely use hate speech again in the future.¹²⁶ Similar efforts may be useful in reaching the so-called moveable middle, such as social media spectators who do not spread hateful content or false information themselves but are open to persuasion in either direction. For example, a study on anti-Roma hate speech in Slovakia found more users left pro-Roma comments on anti-Roma posts after researchers intervened with counter-speech.¹²⁷

Other studies have explored how moral and emotional framings affect audiences, including their perceptions of what is true. Studies of climate change skepticism found that the most effective messages for countering misinformation offer individuals the sense that they

Figure 4. Screenshots of the U.S. Counter-messaging Campaign 'Think Again, Turn Away,' 2014



can take meaningful action, as opposed to messages that portray the world as doomed.¹²⁸ A review of public health messaging found some audience segments were moved more by calls to protect themselves or loved ones than by appeals to social responsibility.¹²⁹

The speaker of the counter-message seems to be quite important. Studies in the rural United States found that friends and family members, community organizations, religious leaders, and medical professionals were the most effective messengers in responding to COVID-19 rumors. In India, health professionals and peers were found to be the most trusted.¹³⁰ Given the influence of informal messengers like social peers, analysts have considered the possibility of using them for official objectives.¹³¹ Volunteer groups countering disinformation, such as the Lithuanian Elves or the North Atlantic Fella Organization, can bring scale,

authenticity, and creativity—traits that official efforts often lack.¹³² Likewise, organic content used to rebut extremist claims and narratives appears more persuasive than government-created content.

There is a risk that poorly designed counter-messaging campaigns can entrench or elevate the very views being rebutted.¹³³ A U.S. Department of State campaign called Think Again, Turn Away illustrates this problem. The anti-Islamic State campaign, launched in 2013, engaged directly with extremists on Twitter but was ultimately deemed counterproductive. Its graphic content and combative tone increased the visibility of Islamic State accounts that replied to the campaign's posts with anti-U.S. rhetoric, while forcing the State Department to engage on unflattering topics like the torture of Iraqi prisoners at the Abu Ghraib prison.¹³⁴ Critics have claimed that Think Again, Turn Away was not focused on the drivers of online extremism and was too clearly affiliated with the U.S. government to serve as a credible messenger. These shortcomings point to the complexities of effective counter-messaging and the need to carefully think through message control, effective messengers, appropriate mediums, and characteristics of the target audience.

HOW EASILY DOES IT SCALE?

Counter-messaging faces implementation challenges due to its often reactive nature. Campaigns frequently arise in response to a belated recognition that disinformation narratives have already grown in strength and impact. Such narratives may have roots going back years, decades, or longer, and their adherents can build up psychological investments over a lifetime. The narratives underpinning disinformation also often evoke powerful emotions, like fear, which can be difficult to defuse once activated.¹³⁵ To mitigate disinformation's first-mover advantages, counter-messengers can try to anticipate such narratives before they spread—for example, predicting attacks on mail-in voting during the 2020 U.S. election—but this is not always feasible.

The need to tailor counter-messaging to a specific audience and context makes scaling more difficult. Reaching large audiences may require breaking them into identifiable subpopulations, each of which would then receive its own research, message development, and novel or even competing strategies. Opting instead for a more generic, large-scale campaign risks undercutting much of the specificity associated with effective counter-messaging. Moreover, broad campaigns increase the odds of misfires, such as the use of messages or messengers that persuade one audience while making another audience double down on its initial beliefs. Elevating rumors or extremist viewpoints is a particular concern. When a

concerning narrative is not yet widespread, campaigners may want to pair strategic silence on the national stage with more discrete messaging that targets specific populations more likely to encounter the narrative.¹³⁶ When the narrative at issue has already become popular, a broad counter-messaging strategy may be appropriate. New digital technologies have the potential to make counter-messaging cheaper and easier to scale, just as innovation can aid in spreading disinformation.

Given the costs of effective counter-messaging at scale, many campaigns seem only modestly funded. The State Department's now-shuttered Center for Strategic Counterterrorism Communications spent only \$6 million on digital outreach in 2012, the year before it launched Think Again, Turn Away.¹³⁷ The center's successor entity, the Global Engagement Center, had a budget of more than \$74 million in 2020.¹³⁸ Australia's COVID-19 vaccine awareness campaign—which included multiple mediums and consultants for outreach to specific vulnerable communities—cost about \$24 million.¹³⁹ For comparison, major brands spend much, much more on advertising (about 10 percent of total revenue, according to one survey).¹⁴⁰ Volunteer-driven efforts, like the North Atlantic Fella Organization, may be appealing partners for external funders due to their low cost and high authenticity. However, overt official support for such activities can diminish their credibility. Extremism scholar Benjamin Lee suggests that looser relationships involving “provision of tools and training” might mitigate this risk.¹⁴¹

CASE STUDY 6

CYBERSECURITY FOR ELECTIONS AND CAMPAIGNS

DESCRIPTION AND USE CASES

Cybersecurity improvements have been proposed as a way to mitigate two distinct kinds of election-related disinformation and influence threats. One threat is hack-and-leak operations, which involve the theft and public exposure of sensitive information about candidates, campaigns, and other political figures. Leaked data may be partially modified or fully authentic. Russian state actors carried out notable hack-and-leaks during the U.S. presidential election in 2016, the French presidential election in 2017, and the UK general election in 2019.¹⁴² To prevent hack-and-leaks, many experts have called for increased cybersecurity protection of candidates, campaigns, and political parties, as well as government offices involved in election processes. This can be done through improved adherence to cybersecurity best practices, donated or discounted cybersecurity services, and specialized training, among other options.¹⁴³ Importantly, such efforts should extend to personal accounts and devices, not just official ones. In 2019, the U.S. Federal Election Commission issued an advisory opinion that some political campaigns could receive free cybersecurity assistance from private firms without violating rules on corporate campaign contributions.¹⁴⁴

The second threat is that hackers may probe or compromise election systems, such as the networks that hold voter registration data or vote tallies. If these operations are discovered and publicized, they can heighten fear that election outcomes are subject to manipulation, thereby reducing confidence in the results—even if this fear is unwarranted. For example, a declassified report by the U.S. Senate Select Committee on Intelligence found that in 2016, Russian actors were in a position to delete or modify voter registration data but did not do so. Other U.S. election infrastructure was probed for vulnerabilities, but there was no evidence to suggest vote totals were modified.¹⁴⁵

KEY TAKEAWAYS:

There is good reason to think that campaign- and election-related cybersecurity can be significantly improved, which would prevent some hack-and-leak operations and fear-inducing breaches of election systems. The cybersecurity field has come to a strong consensus on certain basic practices, many of which remain unimplemented by campaigns and election administrators. Better cybersecurity would be particularly helpful in preventing hack-and-leaks, though candidates will struggle to prioritize cybersecurity given the practical imperatives of campaigning. Election systems themselves can be made substantially more secure at a reasonable cost. However, there is still no guarantee that the public would perceive such systems as secure in the face of rhetorical attacks by losing candidates.

KEY SOURCES:

- “Recommendations to Defend America’s Election Infrastructure,” Brennan Center for Justice, October 23, 2019, <https://www.brennancenter.org/our-work/research-reports/recommendations-defend-americas-election-infrastructure>.
- Erik Brattberg and Tim Maurer, “Russian Election Interference: Europe’s Counter to Fake News and Cyber Attacks,” Carnegie Endowment for International Peace, May 2018, https://carnegieendowment.org/files/CP_333_BrattbergMaurer_Russia_Elections_Interference_FINAL.pdf.
- William Adler and Dhanaraj Thakur, “A Lie Can Travel: Election Disinformation in the United States, Brazil, and France,” Center for Democracy and Technology, December 2021, <https://cdt.org/wp-content/uploads/2021/12/2021-12-13-CDT-KAS-A-Lie-Can-Travel-Election-Disinformation-in-United-States-Brazil-France.pdf>.

The cybersecurity of election systems can often be improved by implementing standard best practices applicable to any organization, such as proactively monitoring network activity, conducting penetration testing, and developing incident response plans. But election systems may also need security measures tailored to their unique context. Such actions can include regularly backing up voter registration databases, certifying voting machines,

maintaining a paper trail for electronic ballots, and conducting post-election audits.¹⁴⁶ The cybersecurity of election systems is intertwined with other aspects of election administration. For example, maintaining accurate electronic tallies of votes depends in part on ensuring that any paper ballots are physically secure and that election workers are properly supervised. (The role of electronic voting machines, a major policy question, is beyond the scope of this report.¹⁴⁷)

Coordination, transparency, and communication are also areas of focus. The U.S. Department of Homeland Security has designated election systems as “critical infrastructure,” allowing it to create structures for better communication between stakeholders and to provide security assistance, such as free cybersecurity assessments for election administrators.¹⁴⁸ Other U.S. examples include the Elections Infrastructure Information Sharing & Analysis Center, a voluntary coordination body created in 2020, and proposals for a national public database of voting system defects.¹⁴⁹

HOW MUCH DO WE KNOW?

The threat of cyber operations against campaigns and election infrastructure is well documented across several countries, but there are few detailed evaluations of the cybersecurity response. In general, the cybersecurity field has come to a strong consensus on certain basic practices to protect against threats. These include multifactor authentication, routine backups (kept segregated from originals), frequent patching, and vulnerability testing. Other actions or principles that have gained favor in recent years include cloud migration, zero trust architecture, and threat intelligence. However, there is very little quantitative evidence of these practices’ comparative cost-effectiveness. Additionally, it is hard to judge the efficacy of best practices in thwarting a highly capable, persistent state actor.

There is a clear causal link between improving campaign cybersecurity and reducing the risk of hack-and-leak operations. With election systems, however, cybersecurity is only half the battle. To maintain public confidence in election integrity, administrators must also *convince* people that systems are truly secure. This critical second step has received less attention from researchers and analysts.

HOW EFFECTIVE DOES IT SEEM?

There is good reason to think that campaign- and election-related cybersecurity can be significantly improved. A 2018 assessment of election administration in all fifty U.S. states found that a distressing number of states had not taken basic precautions, such as minimum cybersecurity standards for voter registration systems.¹⁵⁰ This state of affairs may not be uncommon across government bodies in many countries. A 2022 cybersecurity audit of the

U.S. federal government found that eight of the twenty-three assessed agencies showed significant deficiencies in their ability to detect cyber incidents and protect themselves through basic policies like multifactor authentication and data encryption.

In other words, there are still simple ways to improve cybersecurity in many governmental and political institutions, including campaign and election infrastructure.¹⁵¹ Moreover, such investments would probably prevent a number of intrusions. A 2022 study by the research consultancy ThoughtLab found that organizations which performed well against the National Institute of Standards and Technology (NIST) Cybersecurity Framework, a common benchmark used in many public- and private-sector organizations in the United States and elsewhere, suffered somewhat fewer damaging cyber incidents than lower-performing organizations.¹⁵²

Table 2. U.S. Government Recommendations for Securing Election Systems

Best Practice	Summary
Software and patch management	Create an inventory of software in use by the organization. Deploy patches in a timely manner.
Log management	Maintain secure, centralized logs of devices on and off the network. Review logs to identify, triage, and assess incidents.
Network segmentation	Create separate virtual or physical networks for each part of the organization. Use dedicated systems for election-related tasks.
Block suspicious activity	Enable blocking, not just alerting, of suspicious activity by default. Scan emails and train employees on phishing attacks.
Credential management	Require strong passwords and multi-factor authentication.
Establish a baseline for host and network activity	Track the amount, timing, and destination of typical network traffic to identify anomalies. Create a “gold image” of hosts for comparison.
Organization-wide IT guidance and policies	Maintain incident response and communications plans, an approved software list, and other policies for cyber hygiene.
Notice and consent banners for computer systems	Require that users consent to monitoring, disclosing, and sharing of data for any purpose.

Source: “Best Practices for Securing Election Systems,” U.S. Cybersecurity and Infrastructure Security Agency, November 11, 2022, <https://www.cisa.gov/news-events/news/best-practices-securing-election-systems>.

In addition to prevention, the NIST framework also emphasizes preparedness to respond to and recover from an incident. The 2017 French presidential election provides a celebrated example: Emmanuel Macron's campaign prepared for an eventual Russian hack-and-leak operation by creating fake email addresses, messages, and documents so that stolen materials could not be verified and might discredit the leakers.¹⁵³ Immediate disclosure of all hacking attempts, both to authorities and the public, also built awareness of the disinformation threat to the election. This could be seen as a form of inoculation, or "pre-bunking," which refers to anticipating a specific disinformation narrative or technique and proactively confronting it before it spreads.¹⁵⁴ However, it seems likely that other political, legal, and media factors also played a role in diminishing the influence of the Russian operation.

Unfortunately, public fears of election irregularities cannot always be allayed by truthful assurances that election systems are secure. In United States (in 2020–2021) and Brazil (in 2022–2023), false rhetorical attacks on the integrity of the electoral process by losing candidates and their supporters led to organized postelection violence.¹⁵⁵ A side-by-side comparison of the two examples is revealing, because the two countries have substantially different voting systems. In the United States, a complex set of rules and practices delayed the vote count in a number of states, which laid the groundwork for conspiracy theories of electoral manipulation despite the presence of extensive safeguards and paper-backed auditing mechanisms.¹⁵⁶ Brazil, in contrast, has an all-electronic voting system that allows for rapid results—though it lacks a paper trail to enable physical audits.¹⁵⁷ Despite these divergent approaches, both countries were destabilized by disinformation about election security.

HOW EASILY DOES IT SCALE?

Improving the cybersecurity of political campaigns faces significant cultural and leadership barriers. Campaigns are ephemeral, frenetic environments. They employ large numbers of temporary workers and volunteers who are minimally vetted and trained. Democratic politics also has an inherently open quality—candidates and surrogates must interact with wide swaths of the public, both in person and online—that runs at cross-purposes with physical and cyber security. Finally, a dollar spent on cybersecurity is a dollar not spent on winning votes.

Given these factors, campaigns and candidates often resist making cybersecurity a priority. In the EU, for example, political parties have chronically underfunded their own digital security.¹⁵⁸ A dedicated EU fund could help, but politicians would still need to spend scarce time and attention on cybersecurity and accept the inconveniences that sometimes come with it. When the Netherlands offered cybersecurity training to politicians and government officials before the country's 2017 elections, few expressed interest.¹⁵⁹ One-off or annual trainings are also less effective than more frequent trainings—let alone cultural and orga-

nizational shifts in behavior mandated and enforced by leadership.¹⁶⁰ While some cultural shifts have indeed occurred in recent years across many countries, in campaigns and more generally, political campaigns will likely continue to lag behind other major organizations in their cybersecurity practices.

One advantage of cybersecurity, as compared to other disinformation countermeasures, is that a proven set of best practices already exists and has been widely (if inconsistently) adopted in other sectors. This makes scaling much easier. However, cybersecurity is not necessarily cheap. The size and complexity of national elections and the number of necessary improvements mean that—in the United States, at least—the sums required are significant.¹⁶¹ In 2018, for example, the U.S. Congress allocated \$380 million for election security improvements—including cybersecurity—with millions more given by state governments.¹⁶² And in 2020, the COVID-19 relief bill allocated another \$400 million for elections, with state officials often prioritizing cybersecurity in their grant requests.¹⁶³ Experts tend to propose even larger and more sustained expenditures.¹⁶⁴ The Brennan Center for Justice has called for five-year allocations of \$833 million to help state and local governments with cybersecurity, \$486 million to secure voter registration infrastructure, and \$316 million to protect election agencies from “insider threats.”¹⁶⁵

The cost of securing election infrastructure, while not trivial, seems modest given its foundational importance to democracy.

The cost of securing election infrastructure, while not trivial, seems modest given its foundational importance to democracy. Still, governments must find the political will to make such investments. Proposed measures to improve the security of the 2019 elections for the European Parliament faced resistance from member states that viewed the problem as overhyped or were themselves complicit in election disinformation.¹⁶⁶

CASE STUDY 7

STATECRAFT, DETERRENCE, AND DISRUPTION

DESCRIPTION AND USE CASES

When disinformation and other influence operations stem from abroad, governments can use a range of foreign policy tools to respond. These include sanctions, indictments, media regulation or bans, public statements by government officials, and cyber operations.

The U.S. government has been particularly prolific in many of these areas. After the Russian effort to interfere in the 2016 election, Washington announced a number of sanctions on Russian individuals and organizations.¹⁶⁷ It also announced criminal charges—in 2018 against five Russian organizations and nineteen Russian individuals, in 2021 against two Iranian men, and in 2022 against three Russian men, among others.¹⁶⁸ Although indictments of foreign nationals for influence operation–related activities are unusual globally, sanctions are becoming more common.¹⁶⁹ In 2022, the United Kingdom sanctioned several individuals and media outlets accused of serving as propagandists for Moscow.¹⁷⁰ The same year, a report from the European Parliament noted that the EU does not have a “specific regime of sanctions related to foreign interference and disinformation campaigns orchestrated by foreign state actors” and called for one to be developed.¹⁷¹ Also that year, in the lead-up to the U.S. midterm elections, the U.S. government announced its use of “full-spectrum cyber operations” to “defend and disrupt” influence attempts and “impose costs on foreign actors who seek to undermine democratic processes.”¹⁷²

Overt foreign influence activities can be regulated or banned altogether. The United States has in recent years stepped up enforcement of the Foreign Agents Registration Act, which requires state-affiliated media and others speaking on behalf of foreign interests to disclose

their sponsorship. Similar rules went into effect in 2018 in Australia (where the government also made covert influence efforts a crime) and were proposed in 2022 in the United Kingdom.¹⁷³ Restrictions can also be imposed on foreign state media outlets: the EU and United Kingdom both banned Russian state broadcaster RT following Russia's full-scale invasion of Ukraine.¹⁷⁴ (In the United States, cable providers also dropped the station, leading it to cease operations there.) These bans extended to the outlet's online presence. Some governments have also banned foreign social media platforms seen as vehicles for adversarial influence and espionage: Ukraine banned Russia's VKontakte in 2017 and India banned TikTok, which is owned by a Chinese company, in 2020.

KEY TAKEAWAYS:

Cyber operations targeting foreign influence actors can temporarily frustrate specific foreign operations during sensitive periods, such as elections, but any long-term effect is likely marginal. There is little evidence to show that cyber operations, sanctions, or indictments have achieved strategic deterrence, though some foreign individuals and contract firms may be partially deterrable. Bans on foreign platforms and state media outlets have strong first-order effects (reducing access to them); their second-order consequences include retaliation against democratic media by the targeted state. All in all, the most potent tool of statecraft may be national leaders' preemptive efforts to educate the public. Yet in democracies around the world, domestic disinformation is far more prolific and influential than foreign influence operations.

KEY SOURCES:

- Keir Giles, "Countering Russian Information Operations in the Age of Social Media," Council on Foreign Relations, November 21, 2017, <https://www.cfr.org/report/countering-russian-information-operations-age-social-media>.
- Gabriel Band, "Sanctions as a Surgical Tool Against Online Foreign Influence," Lawfare, September 15, 2022, <https://www.lawfaremedia.org/article/sanctions-surgical-tool-against-online-foreign-influence>.
- Yevgeniy Golovchenko, "Fighting Propaganda with Censorship: A Study of the Ukrainian Ban on Russian Social Media," *The Journal of Politics* 84 (2022), <https://www.journals.uchicago.edu/doi/10.1086/716949>.

Figure 5. Screenshot of an Iranian Disinformation Site Shut Down by U.S. Sanctions, 2020



Finally, many experts have called on political leaders to “be open and outspoken about the nature of the challenge” because “awareness . . . of Russian [and other] information warfare is the most potent defense against it.”¹⁷⁵ This may entail general warnings of the threat: for example, then Finnish defense minister Carl Haglund in 2014 and then British prime minister Theresa May in 2017 both gave blunt descriptions of the overall challenge posed by Russian influence operations.¹⁷⁶ Analysts also suggest using more direct statements tactically during—or, better yet, in advance of—specific influence operations. Experts praised Macron’s campaign for its preparedness against the Russian hack-and-leak operation in 2017, the UK government for its quick response to the poisoning of Sergei and Yulia Skripal in 2018, and the U.S. government for its tactical release of intelligence in 2022 to pre-bunk potential Russian false flag operations in Ukraine.¹⁷⁷

HOW MUCH DO WE KNOW?

There appears to be no empirical scholarship measuring the success of sanctions, indictments, or cyber operations as counter-disinformation tools. Some useful analogies can be drawn from cases where these tools were used against other kinds of hostile activities, such as foreign cyber operations. But this literature, too, is largely anecdotal. Intelligence agencies may be better positioned to assess the efficacy of foreign policy tools by directly observing how foreign actors privately perceive and react to such moves. Such intelligence is mostly kept secret, however, and government officials have not given clear public

characterizations. Cyber operations against disinformation actors offer a qualified exception: U.S. officials have claimed some tactical and operational disruptions. An example is the U.S. cyber operation targeting the Internet Research Agency in Russia prior to the 2018 midterm elections.¹⁷⁸ (Many of these government actions have occurred in tandem with platforms' removals of inauthentic asset networks, discussed in case study 8.)

Bans on both broadcast and digital foreign state media have been subject to a small number of limited empirical studies. The best-studied tool may be proactive and assertive statements by officials, which have a real (though circumstantial) research basis. Although these statements have not received direct empirical study, related areas of scholarship support the notion that they can help prepare and protect the public.

HOW EFFECTIVE DOES IT SEEM?

Foreign policy tools may have varied goals when used against influence operations. They can seek to deter further influence activity (by the same foreign actors or other foreign actors), disrupt foreign influence capabilities (especially during sensitive periods such as elections), or exert a signaling effect (educating the public, affirming redlines to adversaries, and rallying the international community).¹⁷⁹

A broad-based deterrence that leads foreign actors to halt all or much of their influence operations seems out of reach so far, probably because the punishments inflicted to date are not severe enough to outweigh the perceived benefits of influence operation for state perpetrators. For example, despite numerous U.S. sanctions and indictments, Russia and Iran persisted as the largest sources of influence operations removed by Facebook, with many of their operations across platforms targeting the United States.¹⁸⁰ More realistically, democratic states could aim for so-called micro-level deterrence: by targeting low- and mid-level individual perpetrators with sanctions that limit their personal travel, finances, and relationships, these individuals and some others could be dissuaded from participating in influence operations.¹⁸¹ For foreign governments, micro-level deterrence would impose modest organizational costs and friction over time. There is little direct public evidence this works, but it is plausible under the right circumstances.¹⁸²

A concrete example of operational friction came in 2018, when U.S. Cyber Command reportedly disrupted the digital infrastructure of Russia's Internet Research Agency. Officials told the *Washington Post* that "[t]he blockage was so frustrating to the trolls that they complained to their system administrators about the disruption."¹⁸³ U.S. officials offered equivocal assessments of the operation's larger effects. One senator claimed that it successfully prevented "very serious" Russian election interference, while a defense official said the goal was simply to "throw a little curveball, inject a little friction, sow confusion." As Washington continues to institutionalize this kind of activity as a routine part of its federal election plans, the goal of imposing "a little friction" on adversaries for specific time periods seems like a realistic long-term aspiration.¹⁸⁴

For media bans, research on Ukraine's 2017 ban of the Russian social media service VKontakte provides a small window into the potential efficacy of similar tactics elsewhere. A study found that "the sudden censorship policy reduced activity on VKontakte," even among pro-Russia users and despite the ban being "legally and technically" easy to circumvent.¹⁸⁵ This indicates that media bans can have impacts despite imperfect enforcement, because convenience of access remains an important driver of media consumption. Similar dynamics have been observed in other highly restricted media environments, such as China.¹⁸⁶ Likewise, Russian state media channels on U.S. platforms like YouTube, Facebook, and Twitter suffered significant declines in engagement in 2022 due to platforms' efforts to block these pages or limit their reach, though a cat-and-mouse game ensued as Russian outlets created new accounts and channels to evade platform restrictions.¹⁸⁷

These results, if generalizable, would imply that restrictions on foreign media have strong first-order effects. Further research on the VKontakte ban and similar bans should measure not only first-order effects (like activity on the banned platform) but second- and third-order effects as well—to determine, for example, whether pro-Russia users migrated to other platforms, remained interested in pro-Russia propaganda, or lost further trust in the Ukrainian government. It should also examine whether any reciprocal bans of Western media by the Russian government counterbalanced the potential benefits of restrictions on Russian state media.

Leaders' public statements calling attention to foreign influence operations, while not the subject of direct empirical study, have substantial indirect grounding in evidence. Long-standing psychological research holds that when individuals are aware of and plan to resist attempts at persuasion, it makes their original beliefs stronger; analysts today suggest this creates a first-mover advantage through which early warnings of an influence operation can harden resistance to it.¹⁸⁸ A related idea also supported by the literature is pre-bunking.¹⁸⁹ Bipartisan statements may be especially effective: evidence from fact-checking studies shows that corrective measures against disinformation are more effective when the speaker belongs to the same party as the audience (and are less effective when they do not).¹⁹⁰ A barrier to this approach arises when partisan leaders see no advantage in debunking a disinformation narrative, or when they seek political advantages by discrediting the work of fact-checkers. There is also the potential risk of so-called perception hacking, when an operation causes public concern out of line with its actual effect, diminishing public confidence and trust. Officials should be wary of overreactions that play into the hands of disinformers.¹⁹¹

Perhaps the most important limitation on statecraft as a counter-influence tool is the fact that foreign actors are responsible for only a small portion of disinformation. Researchers and investigators around the

The most important limitation on statecraft as a counter-influence tool is the fact that foreign actors are responsible for only a small portion of disinformation.

world overwhelmingly agree on this broad principle—notwithstanding the diverse information environments in different countries and the powerful professional and political incentives to emphasize foreign threats.¹⁹² Domestic actors are almost always more numerous, vocal, resourced, sophisticated, networked, and invested in local political outcomes than foreign entities. This does not mean that foreign influence should be ignored. But it suggests that policymakers' intense focus on the issue may not reflect a realistic assessment of risk.¹⁹³ Rather, foreign disinformation may garner disproportionate attention for the simple reason that many democratic governments have more freedom of action in foreign than domestic policy, especially when it comes to regulating information.

HOW EASILY DOES IT SCALE?

Sanctions, indictments, and media regulations are generally cheap compared to other countermeasures, but they require nontrivial bureaucratic resources to design, develop, and enforce. In the United States, sanctions and indictments targeting foreign influence actors have sometimes been announced many months or even years after the influence activity. Others have been imposed much more quickly. This suggests that the difficulty of investigating and punishing foreign influence actors may depend on circumstantial factors, such as the amount of evidence immediately available, the preexisting intelligence on the responsible actors, and the perceived urgency of a response. Cyber operations are generally more time- and resource-intensive for governments than sanctions, indictments, or media regulations.

Increasing the use of these tools could come at the cost of other foreign policy objectives. Countries may worry that punitive measures will trigger retaliation, jeopardize unrelated negotiations with the targeted state, set diplomatic precedents that constrain their own behavior, or expose intelligence sources and methods, among other possibilities. In 2016, for example, the administration of then U.S. president Barack Obama took limited steps to thwart Russian influence but “did not know the full range of Moscow’s capabilities” and was afraid that harsh measures would prompt even more aggressive election disruption, according to a bipartisan Senate report.¹⁹⁴ The administration also feared domestic confidence in the election’s outcome and fairness would be negatively impacted if it did not tread carefully.¹⁹⁵ Additionally, Washington was trying at the time to coax Russia into helping end the Syrian civil war.

In another case, Western countries’ bans of RT in 2022 prompted Moscow to retaliate by banning the BBC, Voice of America, and other Western services in Russia—restricting their ability to deliver alternative perspectives on the Ukraine war in Ukraine.¹⁹⁶ Moscow and other state adversaries have used this kind of state media restriction to question Western countries’ commitment to free speech abroad and have even used it as a badge of honor by establishing new channels like “¡Censúrame otra vez!” (“Censor me again!”) to target non-Western audiences.¹⁹⁷

CASE STUDY 8

REMOVING INAUTHENTIC ASSET NETWORKS

DESCRIPTION AND USE CASES

Social media companies in the past five years have detected and removed many different networks of fake assets—such as accounts, pages, groups, and events—used to manipulate platforms. These assets are inauthentic, meaning they misrepresent their identity and purpose, and they work together as part of a disinformation campaign or influence operation. Such operations may involve automated activity, human control, or a blend of the two.¹⁹⁸

Major platforms hire in-house investigators, contract out to third-party firms, and form partnerships with researchers to detect and respond to inauthentic asset networks. Platforms also receive tips from governments based on intelligence or law enforcement information. Separately, governments sometimes seek to induce platforms to remove content—inauthentic or otherwise—via legal processes, informal requests, or political pressure.¹⁹⁹ (Removal of authentic content, such as accounts that openly belong to terrorist groups, is beyond the scope of this case study.)

When social media companies remove influence operations asset networks, they often reference objectives like the preservation of authenticity, the right to access reliable information, or the importance of safeguarding elections.²⁰⁰ The policy terms and definitions they use to define the problem differ.²⁰¹ Meta uses the term “coordinated inauthentic behavior” when referring to the use of multiple assets to mislead the company or the public or to evade enforcement of Meta’s policies; this term has become a shorthand for industry ob-

servers and has subsequently been adopted by TikTok (which provides fewer details on its definition).²⁰² X refers to “information operations” in its policies on platform manipulation and spam, and it bans “inauthentic engagements” and “coordinated activity, that attempts to artificially influence conversations through the use of multiple accounts, fake accounts, automation and/or scripting.”²⁰³ Google’s threat analysis group (which also covers YouTube) refers to “coordinated influence operations.”²⁰⁴

KEY TAKEAWAYS:

The detection and removal from platforms of accounts or pages that misrepresent themselves has obvious merit, but its effectiveness is difficult to assess. Fragmentary data—such as unverified company statements, draft platform studies, and U.S. intelligence—suggest that continuous take-downs might be capable of reducing the influence of inauthentic networks and imposing some costs on perpetrators. However, few platforms even claim to have achieved this, and the investments required are considerable. Meanwhile, the threat posed by inauthentic asset networks remains unclear: a handful of empirical studies suggest that such networks, and social media influence operations more generally, may not be very effective at spreading disinformation. These early findings imply that platform takedowns may receive undue attention in public and policymaking discourse.

KEY SOURCES:

- “Threat Report: The State of Influence Operations 2017-2020,” Meta, May 2021, <https://about.fb.com/wp-content/uploads/2021/05/IO-Threat-Report-May-20-2021.pdf>.
- Camille François and evelyn douek, “The Accidental Origins, Underappreciated Limits, and Enduring Promises of Platform Transparency Reporting about Information Operations,” *Journal of Online Trust & Safety* 1 (2021), <https://tsjournal.org/index.php/jots/article/view/17>.
- Gregory Eady et al., “Exposure to the Russian Internet Research Agency Foreign Influence Campaign on Twitter in the 2016 US Election and Its Relationship to Attitudes and Voting Behavior,” *Nature Communications* 14 (2023), <https://www.nature.com/articles/s41467-022-35576-9>.

HOW MUCH DO WE KNOW?

Most information about platforms' takedowns (removals) of inauthentic asset networks comes from the platforms themselves. Major platforms have periodically disclosed takedown actions, sometimes in collaboration with outside investigators. The Disinfodex database contains 326 disclosures by Facebook, Twitter, Google/YouTube, and Reddit between September 2017 and August 2021.²⁰⁵ However, the amount of detail provided—such as the number and type of illicit accounts identified, the narratives employed, the level of user engagement achieved, the perpetrator assessed to be responsible, and the platform policies violated—varies considerably. Among efforts to improve transparency in this area is the EU's 2022 Strengthened Code of Practice on Disinformation, which calls for signatories to standardize data on influence operations across platforms and report them to governments in more granular detail.²⁰⁶

Disclosures of individual takedowns, however informative, still leave open the more fundamental question: To what degree do these takedowns succeed in reducing the prevalence or impact of disinformation? There is virtually no published research that directly bears on this question. For one thing, the success of takedowns depends in part on the responses of bad actors, such as foreign governments and unscrupulous public relations firms. Do takedowns impose enough operational cost and potential public embarrassment on the bad actors to inhibit disinformation campaigns, or do they adjust fairly easily? Platforms and intelligence agencies each have some insight here, but adversarial behavior is inherently hard to observe, may differ from actor to actor, and can change over time.

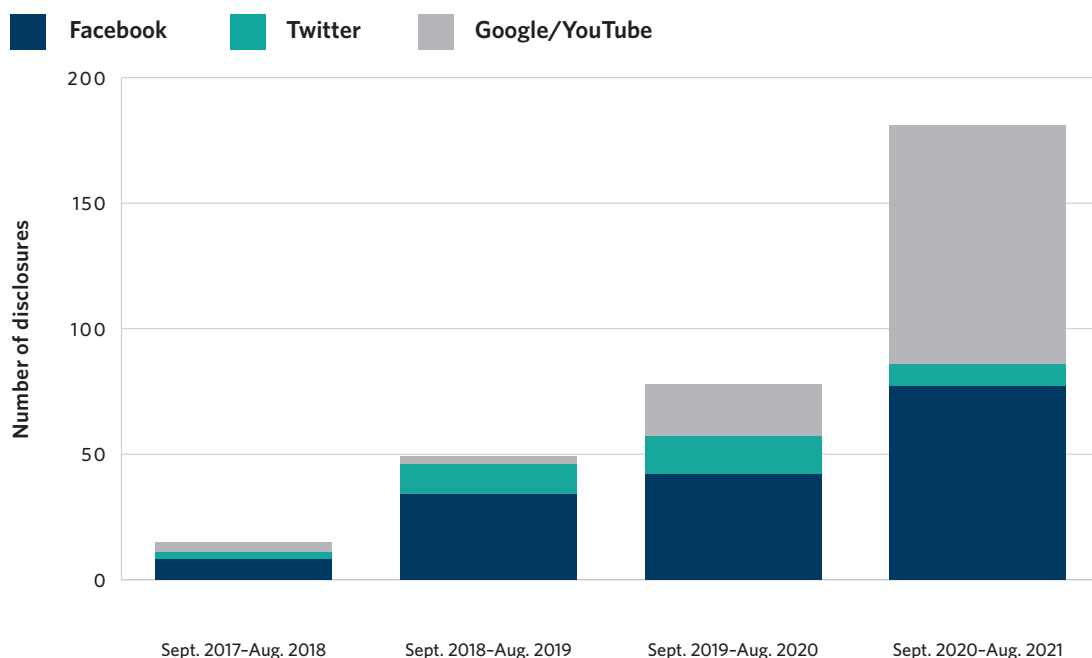
Moreover, platforms do not typically share many (or any) details from their internal research on takedowns, and when fragmentary findings do come to light, they are difficult to interpret in isolation. For example, some platforms have cited positive long-term impacts from their takedowns but did not quantify these claims, specify how they were measured, or share supporting data. Given this gap, leaked copies of internal platform research have provided valuable information to the public. Yet these leaks are infrequent and often come in forms (such as draft, historical, or incomplete documents) or contexts (such as personnel disputes) that partially cloud their meaning. Moreover, very little is known about the effectiveness of social media influence operations themselves and, therefore, how important it is to disrupt them.

HOW EFFECTIVE DOES IT SEEM?

Major social media platforms have greatly increased the number of disclosed takedowns since 2016, which is due in part to a surge of investigative resources. But the covert nature of influence operations makes it hard to say how many remain undetected. If the increase in reported takedowns stems in part from a growing number of influence operations being carried out globally, then the takedowns could be tactically effective but strategically insuf-

ficient. Platforms may also be disclosing a greater portion of the takedowns they perform in response to increased public demands for action. By the same token, news reports of influence operations are increasing year-over-year, which may mean either that operations are increasing or that there is greater public interest in the issue (or both).²⁰⁷

Figure 6. Takedown Disclosures Surged Starting in 2017



Source: Carnegie analysis of “Disinfodex,” Disinfodex, accessed January 9, 2024, <https://disinfodex.org>.

In terms of strategic impact, Facebook reported in 2021 that it had “made progress against [influence operations] by making [them] less effective and by disrupting more campaigns early, before they could build an audience.” Furthermore, these efforts were said to have forced “threat actors to shift their tactics,” pushing them to operate on less popular platforms and invest greater resources in secrecy.²⁰⁸ Such claims are difficult to independently evaluate, however, because the platform provides only limited data access to outside researchers.²⁰⁹

Facebook’s claim of strategic progress appears to be an outlier in the industry; other major platforms have not publicly claimed that their takedowns are having broad effects. This silence may itself be an important indication that broad effects remain difficult to achieve or at least measure. In fact, a draft outside review commissioned by Twitter in 2021 (and leaked by a former executive) found that “analysts are unable to identify and analyze evolving threats or changes in the [tactics, techniques, and procedures] of threat actors,

or measure the effectiveness of action and enforcement, because information is not being preserved.”²¹⁰ Twitter employees apparently believed that the company’s system of “removal ultimately [did] not discourage adversaries from attempting to exploit and leverage the platform, or add costs to their operations because they can quickly adapt.”

The U.S. intelligence community, which has its own unique vantage on foreign influence actors, has offered some tentative praise of platform takedowns. Intelligence analysts looking back at the 2020 election found that Russia’s Internet Research Agency adopted some “short-lived” and perhaps ineffectual new tactics “probably in response to efforts by U.S. companies and law enforcement to shut down” the kind of inauthentic personas used in 2016.²¹¹ Additionally, repeated public disclosures of foreign influence campaigns “probably helped counter them to some degree” in 2020 by improving societal awareness, reducing the deniability of Russia and Iran, and helping China see that covert influence was not “advantageous enough for [it] to risk getting caught meddling.” Nevertheless, U.S. agencies in 2020 “tracked a broader array of foreign actors taking steps to influence U.S. elections than in past election cycles.”

To the extent that social media takedowns can reduce the number and impact of inauthentic asset networks, the ultimate benefit to society depends on the danger posed by those networks. The greater the likely negative effects of an influence operation on society, the more important it is to discover and remove the illicit asset network used to carry out the operation. Unfortunately, very little is known about the effects of online influence operations. Echoing the consensus of experts, Meta’s global threat intelligence lead, Ben Nimmo, calls assessing influence operations’ impact “one of the most difficult questions for investigators.”²¹² (U.S. intelligence agencies are prohibited from even trying to answer the question because it would require analyzing American politics.²¹³)

A Princeton University meta-analysis commissioned by Carnegie found only one rigorous, empirical study of “what has arguably been the greatest focus of policymakers since 2016: the threat of foreign governments using social media to sway voters in democratic elections.”²¹⁴ That study, on Russia’s efforts to influence U.S. Twitter users during the 2016 presidential election, found no effect on user beliefs. Another analysis found that the impact of Russia’s Internet Research Agency in 2016 was likely small.²¹⁵ Meanwhile, a landmark 2018 paper in *Science* suggested that human users are more responsible than automated accounts for the rapid spread of false information online.²¹⁶ Although inauthentic asset networks are not necessarily automated, the finding calls attention to the huge quantity of disinformation being organically generated and disseminated online, unrelated to any

Facebook’s claim of strategic progress appears to be an outlier in the industry; other major platforms have not publicly claimed that their takedowns are having broad effects.

covert influence campaign. These early findings suggest that platform takedowns—while undoubtedly necessary—may receive undue attention in policy conversations about how best to counter disinformation and influence operations.

One limit of removing inauthentic asset networks is that it does not address increasingly prominent coordinated efforts from *authentic* networks to spread disinformation—for example, the online activists who organized the January 6 insurrection in the United States and the January 8 riot in Brazil.²¹⁷ Platforms have fewer rules specifically prohibiting authentic coordinated behavior, even if it results in disinformation, because they are sensitive to accusations of viewpoint censorship and do not want to prevent genuine social movements from coordinating online. This gap in community standards led Facebook to draft a policy against “coordinated social harm,” building on policies against “violence-inducing conspiracy networks,” which the company previously used to justify a ban on the QAnon conspiracy movement.²¹⁸ Recent scholarship suggests that removal of networks like these on one platform (which includes banning or deplatforming authentic user accounts) can have unintended consequences: banned users may migrate to fringe platforms where they create and engage with content more extreme and harmful than what is allowed on mainstream platforms, potentially accelerating their radicalization.²¹⁹

HOW EASILY DOES IT SCALE?

Detecting and removing inauthentic asset networks appears to be moderately expensive, though social media companies have not released specific cost breakdowns. More often, companies give total spending or staffing figures that aggregate many varied aspects of safety and security. As of 2021, Facebook reportedly had “hundreds” of “full-time policy experts tackling foreign influence operations and misinformation”—of which an unknown subset were focused on threat intelligence and asset removal.²²⁰ Twitter/X has had far fewer dedicated staff (the exact number was disputed) even prior to mass layoffs in 2022–2023, and these employees complained internally about antiquated and inadequate tools and technology.²²¹

Comparing the two companies raises challenging questions about what level of investment in takedowns is necessary, cost-effective, and affordable for different platforms. On the one hand, Meta has many times more users than Twitter/X and is a much larger company, so its level of investment should naturally be greater. On the other hand, in countries like the United States, the influence of Twitter/X on political discourse has long been disproportionate to the platform’s overall size.²²² Even so, Meta has had greater ability and incentive to invest in takedowns due to its superior financial performance and the extraordinary public scrutiny it faced after 2016. Zuckerberg told investors in 2017 that “we’re [now] investing so much in security that it will impact our profitability”—a statement of financial strength—whereas Twitter/X has only ever had two profitable years in its history.²²³ The

gap in total safety investments has widened further since businessman Elon Musk bought Twitter and began implementing drastic cost cuts while repudiating many traditional trust and safety practices.

Regardless, it is clear that the further scaling up of takedowns at any platform would require considerable new staffing, technology, and time. According to the 2021 Twitter memo, one employee believed that Twitter could plausibly aim to “add cost to adversaries” (such as those who create inauthentic asset networks) and therefore make strategic progress against them, but that such an effort “would realistically take two years [to] build out.”²²⁴

Additionally, the resources required to investigate illicit asset networks extend beyond the platforms themselves. A growing number of independent companies and nonprofits conduct similar investigations, often in collaboration with platforms. Outside investigators have reported chronic resource constraints and fiscal uncertainty.²²⁵

It is important to note that most major platforms have focused their investigative resources in lucrative Western markets, where advertisers, regulators, and civil society create the greatest pressure for visible action. Independent investigators, too, are heavily concentrated in North America and Europe. There is widespread concern that influence operations continue uninhibited in places where platforms and others have invested fewer investigative resources.²²⁶ This suggests a large untapped opportunity, but it also indicates the high cost of scaling takedowns globally.

CASE STUDY 9

REDUCING DATA COLLECTION AND TARGETED ADS

DESCRIPTION AND USE CASES

Beginning in 2017, two major disclosures—the Cambridge Analytica scandal and the revelations of Russian social media advertising during the 2016 U.S. election—focused Western attention on how personal data can enable political influence.²²⁷ In the ensuing flurry of analysis, disinformation experts began recommending stronger data privacy laws and practices. While these had long been advocated on privacy grounds, they were now also put forth as a means of countering disinformation by reducing the power of microtargeting and algorithmic content moderation.²²⁸

There were at least two related concerns. First, observers feared that microtargeted messages were more effective at mobilizing, persuading, or even manipulating their audiences. Such persuasive power was the central value proposition of social media companies to advertisers, but it also positioned platforms as a potent means of spreading disinformation for both foreign and domestic actors. Today's digital microtargeting represents an evolution of earlier techniques: in the 2004 U.S. presidential campaign, Republican strategist Karl Rove assembled profiles of voters based on personal data, like credit card records, and sent differentiated direct mail ads to each segment.²²⁹ New technology now enables far more advanced methods: online advertisers can use automated tools to develop and deploy an ad in thousands of distinct variations, testing which ones get the most traction with various audiences.²³⁰

KEY TAKEAWAYS:

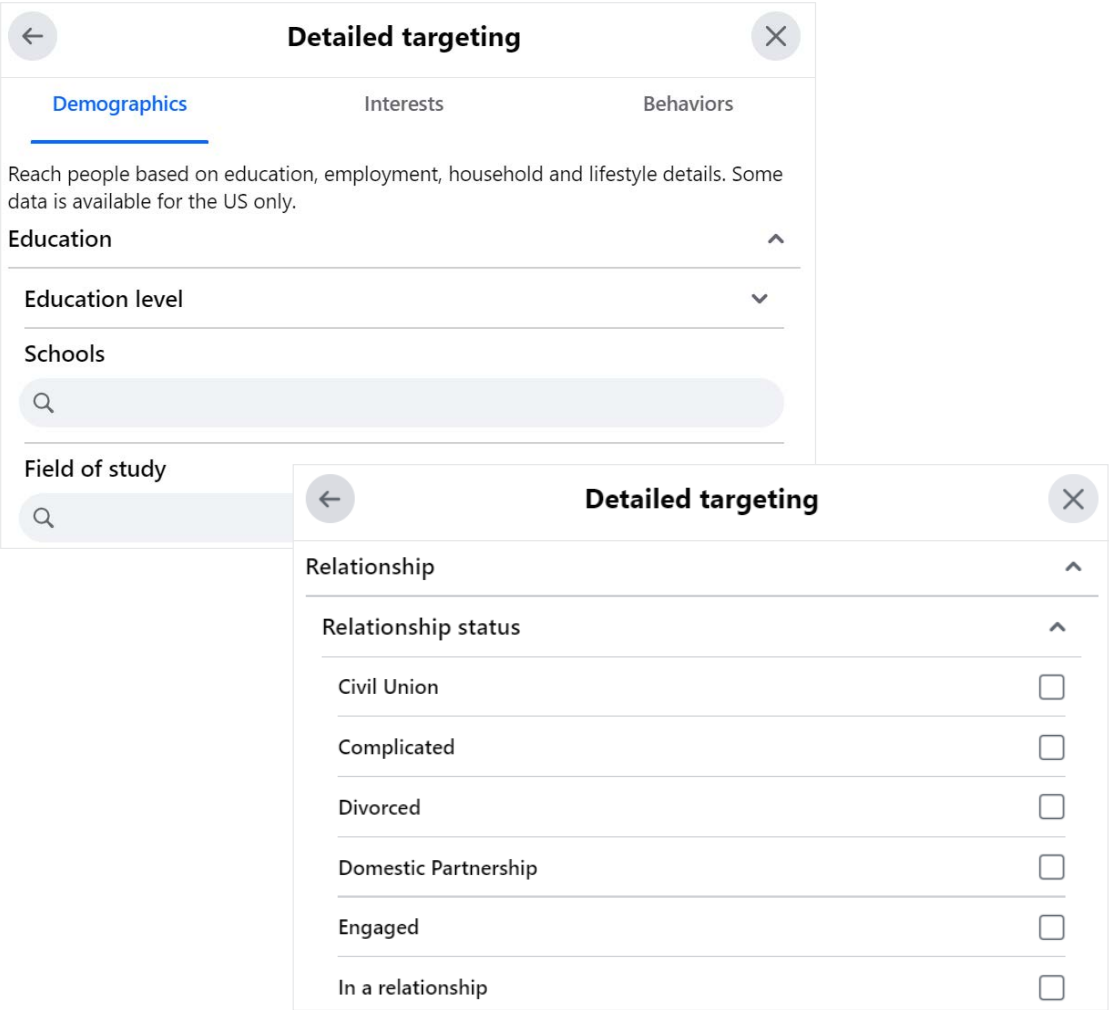
Data privacy protections can be used to reduce the impact of microtargeting, or data-driven personalized messages, as a tool of disinformation. However, nascent scholarship suggests that microtargeting—while modestly effective in political persuasion—falls far short of the manipulative powers often ascribed to it. To the extent that microtargeting works, privacy protections seem to measurably undercut its effectiveness. But this carries high economic costs—not only for tech and ad companies, but also for small and medium businesses that rely on digital advertising. Additionally, efforts to blunt microtargeting can raise the costs of political activity in general, especially for activists and minority groups who lack access to other communication channels.

KEY SOURCES:

- Nathalie Maréchal and Ellery Roberts Biddle, “It’s Not Just the Content, It’s the Business Model: Democracy’s Online Speech Challenge,” *New America*, March 17, 2020, <https://www.newamerica.org/oti/reports/its-not-just-content-its-business-model/>.
- Balázs Bodó, Natali Helberger, and Claes de Vreese, “Political Microtargeting: A Manchurian Candidate or Just a Dark Horse?” *Internet Policy Review* 6 (2017), <https://policyreview.info/articles/analysis/political-micro-targeting-manchurian-candidate-or-just-dark-horse>.
- Dipayan Ghosh and Ben Scott, “Digital Deceit: The Technologies Behind Precision Propaganda on the Internet,” *New America*, January 23, 2018, <https://www.newamerica.org/pit/policy-papers/digitaldeceit/>.

Second, the prevailing business model of social media—use of personal data to curate content in ways that capture users’ attention and serve them more, and more effective, advertisements—also came under criticism. The model was seen as an engine for the spread of misleading, incendiary content, including disinformation, and therefore bore responsibility for serious harm to democracies around the world.²³¹ (See case study 10 for more on social media algorithms.)

Figure 7. Screenshots of Targeting Options for Facebook Advertisers



The largest and highest-profile policy response was the EU’s General Data Protection Regulation (GDPR).²³² The GDPR requires that commercial entities operating in the EU or “monitoring the behaviour of individuals in the EU” receive user consent to collect personal data and that they collect the minimum necessary data for their purpose.²³³ Anonymized data is excluded from the regulation, and organizations must take steps to protect identity through pseudonymization processes like encryption as soon as feasible in

the collection process. Larger companies must have a data protection officer charged with overseeing compliance; violations can incur large fines.²³⁴ Since the GDPR's passage, many policymakers, analysts, and activists have called for the United States and other countries to adopt more privacy policies of their own.²³⁵

The GDPR faced obstacles in limiting the use of personal data for political messaging; for example, some EU member states used their regulatory autonomy to make crucial exemptions for political parties, and many regulations are unclear about which entities are responsible for compliance and what counts as a political advertisement as opposed to an issue advertisement.²³⁶ Some of these gaps are being filled by the EU's Digital Services Act (DSA), which passed in 2022, and further political advertisement rules currently being finalized.²³⁷ They define political advertisements more clearly as ads run by politicians or campaigns or about legislation or elections. They also limit what type of data can be used for political microtargeting; only very basic information like age and language can be used during election periods, and there are permanent restrictions on the use of sensitive data like political orientation.²³⁸ These strengthened restrictions are weaker than some officials proposed. In January 2022, for instance, the EU's data protection agency called for a ban on all microtargeted political ads, as have the European Parliamentary delegations from several countries.²³⁹

Technology companies have also taken voluntary steps to reduce the granularity of microtargeting and the invasiveness of data collection. Perhaps most dramatically, in October 2019, Twitter banned all political advertising—though the company struggled to differentiate political and issue advertisements and reversed this policy after its acquisition by Musk.²⁴⁰ In 2019, Google limited political ad targeting to “age, gender, and general location”; in 2021, the company also announced several new limits on how it would profile and target third-party ads across the internet.²⁴¹ In 2021, Facebook also restricted microtargeting options for political and sensitive topics.²⁴² Additionally, in early 2022, Apple made a change to its iOS software that prevented companies from tracking user activity across third-party apps and websites.

HOW MUCH DO WE KNOW?

Unlike some other types of countermeasures, data privacy has recent achievements that can be evaluated on real-world impact. Unfortunately, while analysts generally find that recent privacy efforts by the EU, tech companies, and others have somewhat reduced the amount and power of microtargeting, there has been very little scholarship that directly assesses their impact on disinformation.²⁴³ Much more analysis has focused on other impacts, such as the economic costs of privacy efforts, which may nevertheless have an indirect relationship to disinformation.²⁴⁴

Additionally, the efficacy of this intervention has much to do with how persuasive data-driven advertising is to begin with. Digital microtargeting's immense power is a key claim of technology companies, advertising firms, and political consultants. The sheer size of the digital ad industry attests to widespread belief in this marketing claim: \$600 billion is spent annually in the sector, which has given rise to several of the world's most valuable companies. However, there is little scholarship or independent analysis to substantiate industry claims, particularly in the context of political advertising.²⁴⁵ A 2020 study noted that "literature on the effects of [political microtargeting] techniques is scarce."²⁴⁶ The dangers of political microtargeting, while often discussed, are still largely unknown.²⁴⁷

Recent studies provide some experimental evidence on the effects of political and commercial microtargeting.²⁴⁸ However, much of the literature on political advertising comes from the United States, which has several distinctive election dynamics—including a polarized two-party system, a loose set of campaign finance regulations, and a remarkably long election season (allowing more advertising of all kinds).²⁴⁹ Research findings from the United States might not apply elsewhere.

Moreover, studies of political microtargeting have generally focused on whether such advertisements change people's votes. Swaying voter preferences, while undoubtedly important, is not always the only or most important effect or purpose of political disinformation. Influencing voter turnout—by energizing supporters, discouraging opponents, or tricking people about voting rules—can be equally or more impactful in some elections. Microtargeted disinformation might also aim to alter political dynamics in more complicated and indirect ways, like by increasing levels of distrust, fear, or apathy. No studies specifically assessing such effects were found.

HOW EFFECTIVE DOES IT SEEM?

Nascent scholarship suggests that microtargeting is somewhat effective for political campaigns but falls far short of the manipulative powers often ascribed to it. A 2018 experiment in the Netherlands examined issue-based targeting—a tactic used by candidates, parties, and other interest groups to show ads about specific topics, such as crime or reproductive rights, to the voters most likely to care about those topics. The study found that ads microtargeted by issue did not increase the salience of that issue but did modestly improve the likelihood that subjects would vote for the political party whose branding was used in the message.²⁵⁰ The authors stressed that microtargeted ads competed with many other sources of information that voters encounter, adulterating the ads' influence. For this reason, microtargeting—like other forms of advertising—may be more effective in smaller elections, like local or primary races, that attract less attention than national contests.²⁵¹

Similar results were found in the few studies of personality-based (also called psychometric or psychographic) microtargeting. These techniques, famously practiced by Cambridge Analytica, seek to target individuals prone to certain feelings, such as anger.²⁵² While Cambridge Analytica is now widely considered to have oversold its capabilities, academic experiments have found that personality-based microtargeted political ads can be more effective than regular ads at increasing positive feelings toward a candidate. However, evidence is mixed about whether such microtargeting does any more than traditional ads to influence viewers' voting preferences—a crucial test of electoral persuasion.²⁵³

If microtargeting is modestly effective, then how much do privacy protections reduce its persuasive power? Evidence suggests the change can be meaningful. One study associated the GDPR with a 12 percent decrease in page views and e-commerce revenue following its implementation.²⁵⁴ Another study estimated that affected businesses suffered an 8 percent loss of profits and a 2.2 percent drop in sales.²⁵⁵ While these studies did not directly examine disinformation or political persuasion, they provide indirect evidence that privacy protections can make digital ads measurably less effective.

HOW EASILY DOES IT SCALE?

Since the EU passed the GDPR, several countries and some U.S. states have implemented their own privacy laws.²⁵⁶ It seems likely that, in time, most democracies will have significantly more privacy protections than they do today. Moves by major tech companies, such as Apple and Google, also suggest a perceived market demand for better privacy.

Privacy has its own inherent and practical value, which should be a major (perhaps primary) factor in any cost-benefit analysis of expanded data protections. On the other hand, studies have shown high compliance costs and even higher lost revenue for firms and advertisers.²⁵⁷ The GDPR, according to some estimates, cost companies hundreds of billions of dollars.²⁵⁸ Similarly, Apple's iOS changes cost Meta an estimated \$10 billion a year in revenue and contributed to a \$250 billion decline in Meta's market value.²⁵⁹ Research by the Information Technology and Innovation Foundation, a pro-business think tank, estimates that stricter U.S. privacy laws could cost the American economy \$122 billion annually.²⁶⁰ Some scholars argue the cost will be lower if many users opt out of the protections—but this will also dilute the law's counter-disinformation potential.²⁶¹ Although giant tech and advertising companies can most likely bear the cost of privacy regulations, much of the cost would fall on others—such as small- and medium-sized enterprises—that rely on digital advertising to attract customers.²⁶²

Rather than reducing data collection across the board, policymakers could instead place stricter limits on the use of microtargeting for political purposes. The EU's DSA takes this approach (albeit on top of existing GDPR rules). However, strict limits on political advertising make it harder for politicians to reach voters. If applied to issue ads, limits can also reduce the reach of activists, who have come to rely on digital advertising to compete with powerful interests and raise the salience of neglected issues or perspectives.²⁶³ Conversely, if restrictions on issue ads are too lenient, such ads become an easy way to avoid restrictions on political microtargeting. One way to address these problems would be to pair data privacy and microtargeting restrictions with broader campaign finance reforms, such as public financing of elections or stronger political spending disclosures.

One way to address disinformation would be to pair data privacy and microtargeting restrictions with broader campaign finance reforms.

CASE STUDY 10

CHANGING RECOMMENDATION ALGORITHMS

DESCRIPTION AND USE CASES

Social media companies employ machine learning algorithms to recommend content to users through curated feeds or other delivery mechanisms. A separate set of algorithms is trained to detect undesirable content, which is then penalized by recommendation algorithms (what some scholars refer to as “reduction”).²⁶⁴ These two processes work together to shape what individuals see on social media.

A major concern is that algorithms contribute to the spread of disinformation on social media because they typically reward engaging content.²⁶⁵ The more users are engaged, the more ad and subscription revenue a company can earn—and unfortunately, disinformation tends to be highly engaging. For instance, a landmark 2018 paper in *Science* found that human users spread false information more quickly than true information.²⁶⁶ Disinformation tends to be not only misleading but also sensationalist, divisive, and laden with negative emotions—seemingly tailor-made for distribution by engagement-seeking algorithms.²⁶⁷ In 2018, internal Facebook research found that the company’s recommendation algorithm promoted emotionally volatile content.²⁶⁸

Political figures and a range of other actors have opportunistically seized on this dynamic. The term “fake news” was popularized during the 2016 U.S. election, when Macedonians and others responded to the financial incentives of this attention economy by generating viral false news stories for U.S. audiences.²⁶⁹ Scholars and journalists have documented how clickbait pages and public relations firms in other countries use similar strategies to drive

engagement and revenue.²⁷⁰ Changes to platform algorithms, though difficult to study, could potentially alter financial and political incentives in ways that constrain the spread of disinformation and increase social resilience to it by reducing polarization. Further, by stopping short of outright removal, reduction can preserve greater freedom of speech while still limiting the impact of disinformation.²⁷¹

KEY TAKEAWAYS:

Although platforms are neither the sole sources of disinformation nor the main causes of political polarization, there is strong evidence that social media algorithms intensify and entrench these off-platform dynamics. Algorithmic changes therefore have the potential to ameliorate the problem; however, this has not been directly studied by independent researchers, and the market viability of such changes is uncertain. Major platforms' optimizing for something other than engagement would undercut the core business model that enabled them to reach their current size. Users could opt in to healthier algorithms via middleware or civically minded alternative platforms, but most people probably would not. Additionally, algorithms are blunt and opaque tools: using them to curb disinformation would also suppress some legitimate content.

KEY SOURCES:

- Tarleton Gillespie, "Do Not Recommend? Reduction as a Form of Content Moderation," *Social Media + Society* 8 (2022): <https://journals.sagepub.com/doi/full/10.1177/20563051221117552>.
- Paul Barrett, Justin Hendrix, and J. Grant Sims, "Fueling the Fire: How Social Media Intensifies U.S. Political Polarization — And What Can Be Done About It," NYU Stern Center for Business and Human Rights, September 13, 2021, <https://bhr.stern.nyu.edu/polarization-report-page?ga=2.126094349.1087885125.1705371436-402766718.1705371436>.
- Joshua A. Tucker et al., "Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature," Hewlett Foundation, March 2018, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3144139.

HOW MUCH DO WE KNOW?

Very few published studies have directly assessed algorithmic adjustments as a counter-disinformation strategy. However, there is a strong circumstantial case—based on related social science, leaked platform research, and theoretical inferences—that platforms’ existing algorithms have amplified disinformation. It is therefore plausible that changes to these same algorithms could ameliorate the problem.

More research is needed on the extent to which new algorithmic changes can help to slow, arrest, or reverse cycles of disinformation and polarization that are already deeply rooted in some democracies.²⁷² For example, would an algorithmic reduction in divisive content cause some users to migrate to other, more sensational platforms? The answers may change as the social media marketplace continues to evolve. The U.S. market has been dominated for several years by a handful of companies that publicly (albeit imperfectly) embrace a responsibility to moderate their platforms. But major platforms’ cuts in trust and safety budgets, Twitter’s ownership change, and the rise of numerous alternative platforms suggest this moment may be passing.

HOW EFFECTIVE DOES IT SEEM?

To be sure, platforms are neither the sole sources of disinformation nor “the original or main cause of rising U.S. political polarization” and similar patterns elsewhere.²⁷³ Moreover, what people view on platforms is not just a function of recommendation algorithms. One 2020 study suggested that individuals’ conscious choices to subscribe to YouTube channels or navigate to them via off-site links were a bigger driver of views on extremist videos than algorithmic rabbit holes, for example.²⁷⁴

Nevertheless, there is good reason to believe that social media algorithms intensify and entrench on- and off-platform dynamics which foment disinformation. A 2018 literature review published by the Hewlett Foundation described a self-reinforcing cycle involving social media, traditional media, and political elites.²⁷⁵ It found that when algorithms amplify misleading and divisive content online, political elites and traditional media have greater incentive to generate disinformation and act and communicate in polarizing ways—because the subsequent social media attention can help them earn support and money. In other words, a combination of online and offline dynamics leads to a degradation of the political-informational ecosystem.

If curation algorithms help to foment disinformation, then changes to those algorithms could potentially help to ameliorate the problem, or at least lessen platforms’ contributions to it. Such changes could fall into at least two broad categories. First, platforms could

When algorithms amplify misleading and divisive content online, political elites and traditional media have greater incentive to generate disinformation and act and communicate in polarizing ways.

machine learning to demote content more aggressively based on the probability it could incite violence or violate other policies.²⁷⁷ The use of machine learning to assess content in this way is called “classification.”²⁷⁸ It allows platforms to moderate content at scale more efficiently than human review, but it is far from an exact science. Classifiers differ greatly in their precision, so any enhanced reliance on reduction strategies might require improving classifier reliability. Human-made rules and judgments are also still an important factor in reduction strategies. Platform staff must decide what types of content algorithms should search for and weigh the risks of penalizing permissible speech or permitting harmful content to go unmoderated. What level of reliability and statistical confidence should a classification algorithm demonstrate before its judgments are used to demote content? Should algorithms only demote content that might violate explicit terms of service, or should other concepts, such as journalistic quality, play a role? Answers to these questions have important implications for freedom of speech online. Regardless, stronger use of classifiers to demote sensational content could plausibly reduce the spread of disinformation.

Recommendation algorithms could also be made to prioritize values, or “curatorial norms,” other than engagement.²⁷⁹ While the major platforms fixate on maximizing engagement because of their commercial interests, a few smaller platforms design user experiences to promote constructive debate and consensus-building—for example, by requiring users to earn access to features (including private messaging and hosting chat rooms) through good behavior.²⁸⁰ Recommendation algorithms can be used for a similar purpose: a 2023 paper explored how algorithms can “increase mutual understanding and trust across divides, creating space for productive conflict, deliberation, or cooperation.”²⁸¹ Similarly, others have advocated for drawing on library sciences to create recommendation algorithms that amplify the spread of authoritative content instead of eye-catching but potentially untrustworthy content.²⁸²

Another idea is to give users more control over what content they see. Several large platforms now allow users to opt in to a “chronological feed” that lists all followed content in order, with no algorithmic ranking. Others have proposed enabling users to pick their own

expand their use of algorithms to reduce the prevalence of certain kinds of content through detection and demotion or removal. Second, they could train algorithms to recommend content in pursuit of values other than, or in addition to, engagement.²⁷⁶

Facebook temporarily used the first strategy during the 2020 U.S. election, employing

algorithms. This could be done through middleware, or software that serves as an intermediary between two other applications—in this case, between the user interface and the larger social media platform, allowing users to choose from competing versions of algorithmic recommendation.²⁸³ It is difficult to evaluate middleware as a counter-disinformation intervention because it has not been widely implemented. Critics say the recommendation algorithm that many users prefer may well be the most sensationalist and hyperpartisan, not the most measured and constructive.²⁸⁴ Political scientist Francis Fukuyama, perhaps the most prominent advocate for middleware, has said that the goal of middleware is not to mitigate misinformation but to dilute the power of large technology companies in the public square.²⁸⁵

Finally, it should be noted that some users are unusually motivated to seek out and consume extreme, divisive, and inflammatory content. Scholarship holds that these users represent a small fraction of users overall, but they can have outsized political impact, both online and offline.²⁸⁶ These atypical users will likely find their way to the content they desire—through searches, subscriptions, messages with other users, and other means.²⁸⁷

HOW EASILY DOES IT SCALE?

The market viability of algorithmic changes is uncertain. Today's large platforms grew to their current size by combining data collection, targeted advertising, and personalized systems for distributing compelling—if often troublesome—content. Dramatic departures from this model would have unknown business consequences, but analogous changes to user privacy suggest it could be in the tens of billions of dollars. When Apple made it more difficult to serve targeted advertisements on iOS devices, Meta claimed it cost the company \$10 billion in annual revenue. Analysts estimated in 2023 that an EU legal ruling requiring Meta to let users opt out of targeted ads could cost the company between 5 and 7 percent of its advertising revenue, which was \$118 billion in 2021.²⁸⁸

It is also unclear that civically oriented platforms, absent public support or heavy-handed regulation, can compete with social media companies which carefully calibrate their services to maximize engagement. Even some proponents of civically oriented platforms have suggested that these are unlikely to scale as well as today's largest platforms because they often serve communities with specific needs, such as those seeking anonymity or highly attentive content moderation. Such platforms may lack key features many users desire, like the ability to share photos or reply directly to posts.²⁸⁹ Regulations can help level the playing field: for example, the DSA now requires the biggest platforms to provide EU users with a chronological feed option.²⁹⁰

Finally, new algorithms intended to reduce disinformation could be just as opaque as today's algorithms, while also having unintended consequences for online speech. Removing content through takedowns or other means is relatively apparent to the public, but algorithmic reduction requires complex, back-end changes that are more difficult for external analysts to detect or study. Additionally, efforts to curb the spread of sensationalist or emotionally manipulative content could also sap the online creativity that makes the internet a driver of culture and commerce. Using algorithms to reduce disinformation, and only disinformation, may well be impossible.

LOOKING AHEAD: GENERATIVE AI

During the course of this research project, rapid advances in generative AI—which can create new content—have led to heated speculation about the future disinformation landscape. AI experts, commentators, and politicians have predicted that generative AI will dramatically worsen the disinformation problem. Many disinformation experts have been more circumspect, but some are also worried. It is too soon to make strong predictions, and this report has primarily focused on assessing countermeasures rather than handicapping emergent threats. Still, some initial possibilities can be laid out, drawing on relevant parallels from this report’s case studies and a review of disinformation research.

Generative AI algorithms, developed with machine learning techniques, can be used to produce a wide variety of text, images, video, and audio. Some results are strikingly similar to authentic, human-produced media—though, at least for now, close observation can often reveal odd artifacts and incongruities. Generative algorithms can facilitate disinformation in countless ways.²⁹¹ Deepfake videos and audio clips have been used to simulate specific individuals saying and doing things they never did, allowing perpetrators to defame or degrade their targets. AI-generated photos of nonexistent people have been incorporated into fake social media profiles to bolster their apparent authenticity. Synthetic text generation can automate the mass production of written propaganda, fraudulent documents, or online conversation in many languages and genres. Additionally, it’s becoming easier to combine multiple generative AI techniques—for example, automating the writing of a film script that is then acted out on video by a lifelike synthetic avatar.

As these examples illustrate, generative AI has several concerning qualities relevant to disinformation. It can enable the rapid, low-cost production of false or misleading content. It can lower the barriers to entry for creating such content—for example, obviating the need for subject matter knowledge, charisma, or skills in video production, photo editing, and translation. AI-generated content can also be more realistic than what its creators could otherwise produce. Additionally, there is worry that algorithms will enable the production of tailor-made content that targets a specific audience by using personal information to predict what will resonate. Finally, generative AI can quickly reformulate preexisting content—be it a terrorist’s manifesto or malicious computer code—in ways that may stymie detection and response.

But the theoretical dangers of new technology do not always manifest as initially feared. Deepfakes, for example, have generated much consternation since 2017, yet they’ve had hardly any political impact so far. To be sure, deepfake-driven nonconsensual pornography is a real problem, and deepfakes are occasionally used to commit fraud. Moreover, technological improvements may make deepfakes harder to spot in the future. Still, it is noteworthy that the most dreaded type of deepfakes—mass political disinformation—has rarely been attempted and has largely failed. Two recent prominent examples were videos of Ukrainian President Volodymyr Zelenskyy and Russian President Vladimir Putin disseminated in their respective countries during the war. Each was carefully orchestrated and timed for maximum political effect.²⁹² They were even broadcast on TV news networks that had been compromised by hackers, enhancing the verisimilitude. Yet neither video seemed to have real persuasive power, perhaps because the target audiences placed greater trust in the national authorities who debunked them. For now, bad actors seem to recognize that cruder techniques—including so-called cheapfakes that employ traditional video manipulation or involve simply relabeling an old photo to change its apparent meaning—remain easier and adequately effective.

The limited impact of political deepfakes so far is consistent with research on disinformation generally and has implications for other kinds of generative AI. Studies suggest that people’s willingness to believe false (or true) information is often not primarily driven by the content’s level of realism. Rather, other factors such as repetition, narrative appeal, perceived authority, group identification, and the viewer’s state of mind can matter more. This has relevance for emergent policy efforts to counter AI-generated disinformation.

People’s willingness to believe false (or true) information is often not primarily driven by the content’s level of realism.

For example, the Coalition for Content Provenance and Authenticity offers a digital “signing authority” that certifies the source of content and allows end users to verify its authenticity.²⁹³ Ultimately, though, such badges of truth depend on people choosing to trust and value them. Many forms of

disinformation thrive today because their believers choose to discount existing sources of high-quality information—such as rigorous professional journalism, enduring scientific consensus, and transparent independent investigations.

The prediction that personalized AI content will enable extremely persuasive disinformation should also be treated with caution. Personalized persuasion is already an enormous industry: commercial and political advertisers spend hundreds of billions of dollars each year trying to reach audiences with content tailor-made for specific demographics, interests, and temperaments. Yet studies of microtargeting have been less than impressive, casting doubt on the notion that data-driven personalization is uniquely compelling. Perhaps generative AI will invent entirely new forms of microtargeting that are much more compelling than anything humans have come up with. But humans have already deployed similar tools—such as big data, applied statistics, and previous forms of AI—for years without any apparent revolutionary breakthroughs in the science of persuasion. As this report’s case studies show, persuasive content—however personalized—must still compete with the cacophony of other influences on targets. Someone’s beliefs are often shaped by a lifetime of narrative and psychological commitments, which are not easily dislodged.

Finally, generative AI is a double-edged sword: it can be used to counter disinformation as well as foment it. Generative AI products like ChatGPT are marketed as general-use tools for information discovery, categorization, and reasoning. If that promise holds, then counter-disinformation practitioners can use generative AI to improve their overall productivity in countless ways. For example, future AI tools—if responsibly used, supervised, and verified by humans—may help fact-checkers more quickly triage and categorize claims, find relevant truthful information, and compare the former with the latter. Such tools could also facilitate the expanded use of social media labels—working in tandem with humans to support faster judgments on source quality, which could then enable platforms to add more decisive labels to more content. AI has obvious applications for cybersecurity and is already a standard feature of many products and services. Algorithms also help social media companies—which possess world-class AI capabilities—identify inauthentic asset networks.²⁹⁴ While content curation algorithms are notorious for contributing to the spread of disinformation, the same engineering principles can be used to design more responsible algorithms. In short, generative AI and other machine learning technologies can be applied to many of the countermeasures explored in this report. In several cases, AI offers promising opportunities to address important cost and scaling challenges.

It is impossible to know the long-term impacts of generative AI. Like other new technologies, it will be employed by all sides of the information contest. If generative AI lives up to its current hype, it could mark the latest in a long series of what Carnegie’s Alicia Wanless calls “information disturbances”—historical developments that alter the information ecosystem, touching off a chaotic cycle of reaction and counterreaction which unsettles, at least

for a time, whatever balance existed before.²⁹⁵ Future historians may debate whether the rise of generative AI turned out to be uniquely disturbing, or just another continuation of a long digital revolution.

Regardless, the metaphor of an “information ecosystem” has a larger lesson for policymakers. The flow of information through society is extraordinarily complex, much like the many forms of competition and cooperation found in the natural world. People’s beliefs, expressions, and actions are shaped by countless psychosocial factors that interact in still mysterious ways. Such factors are rarely, if ever, reducible to any single technology.

Policymakers concerned about disinformation should embrace complexity and acknowledge uncertainty. The effort to counter disinformation will be a long journey through a dark thicket, with many wrong turns and pitfalls along the way. Yet democracies have no choice but to undertake this difficult journey—hopefully guided by the light of evidence, no matter how dim this light may be.

NOTES

- 1 The cells of this table are color coded: green suggests the most positive assessment for each factor, while red is the least positive and yellow is in between. These overall ratings are a combination of various subfactors, which may be in tension: for example, an intervention can be highly effective but only for a short time or with high risk of second-order consequences.
- A green cell means an intervention is well studied, likely to be effective, or easy to implement. For the first column, this means there is a large body of literature on the topic. While it may not conclusively answer every relevant question, it provides strong indicators of effectiveness, cost, and related factors. For the second column, a green cell suggests that an intervention can be highly effective at addressing the problem in a lasting way at a relatively low level of risk. For the third column, a green cell means that the intervention can quickly make a large impact at relatively low cost and without major obstacles to successful implementation.
- A yellow cell indicates an intervention is less well studied (there is relevant literature but major questions about efficacy are unanswered or significantly underexplored), less efficacious (its impact is noteworthy but limited in size or duration, or it carries some risk of blowback), or faces nonnegligible hurdles to implementation, such as cost, technical barriers, or political opposition.
- A red cell indicates that an intervention is poorly understood, with little literature offering guidance on key questions; that it is low impact, has only narrow use cases, or has significant second-order consequences; or that it requires an especially high investment of resources or political capital to implement or scale.
- 2 See “Final Report: Commission on Information Disorder,” Aspen Institute, November 2021, https://www.aspeninstitute.org/wp-content/uploads/2021/11/Aspen-Institute_Commission-on-Information-Disorder_Final-Report.pdf; Daniel Arnaudo et al., “Combating Information Manipulation: A Playbook for Elections and Beyond,” National Democratic Institute, International Republican Institute, and Stanford Internet Observatory, September 2021,

<https://www.ndi.org/sites/default/files/InfoManip%20Playbook%20updated%20FINAL.pdf>; “Center of Excellence on Democracy, Human Rights, and Governance: Disinformation Primer,” U.S. Agency for International Development, February 2021, <https://cnxus.org/wp-content/uploads/2021/11/usaaid-disinformation-primer.pdf>; and Laura Courchesne, Julia Ilhardt, and Jacob N. Shapiro, “Review of Social Science Research on the Impact of Countermeasures Against Influence Operations,” *Harvard Kennedy School Misinformation Review* 2, no. 5 (September 2021): <https://misinforeview.hks.harvard.edu/article/review-of-social-science-research-on-the-impact-of-countermeasures-against-influence-operations/>.

- 3 This list was drawn from multiple sources, including Kamyā Yadav, “Countering Influence Operations: A Review of Policy Proposals Since 2016,” Carnegie Endowment for International Peace, November 30, 2020, <https://carnegieendowment.org/2020/11/30/countering-influence-operations-review-of-policy-proposals-since-2016-pub-83333>; a more detailed, unpublished database of policy proposals compiled by Vishnu Kannan in 2022; Courchesne, Ilhardt, and Shapiro, “Review of Social Science Research”; and “The 2022 Code of Practice on Disinformation,” European Commission, accessed January 27, 2023, <https://digital-strategy.ec.europa.eu/en/policies/code-practice-disinformation>. These sources were supplemented by further literature review and expert feedback.
- 4 Alicia Wanless and James Pamment, “How Do You Define a Problem Like Influence?,” Carnegie Endowment for International Peace, December 30 2019, <https://carnegieendowment.org/2019/12/30/how-do-you-define-problem-like-influence-pub-80716>. For more on the distinction between misinformation and disinformation, see Dean Jackson, “Issue Brief: Distinguishing Disinformation From Propaganda, Misinformation, and ‘Fake News,’” National Endowment for Democracy, October 17, 2017, <https://www.ned.org/issue-brief-distinguishing-disinformation-from-propaganda-misinformation-and-fake-news/>.
- 5 Jim Clapper et al., “Public Statement on the Hunter Biden Emails,” *Politico*, October 19, 2020, <https://www.politico.com/f/?id=00000175-4393-d7aa-af77-579f9b330000>.
- 6 Luke Broadwater, “Officials Who Cast Doubt on Hunter Biden Laptop Face Questions,” *New York Times*, May 16, 2023, <https://www.nytimes.com/2023/05/16/us/politics/republicans-hunter-biden-laptop.html>.
- 7 See, for example, Joseph Bernstein, “Bad News: Selling the Story of Disinformation,” *Harper’s*, 2021, <https://harpers.org/archive/2021/09/bad-news-selling-the-story-of-disinformation>; Rachel Kuo and Alice Marwick, “Critical Disinformation Studies: History, Power, and Politics,” *Harvard Kennedy School Misinformation Review*, August 12, 2021, <https://misinforeview.hks.harvard.edu/article/critical-disinformation-studies-history-power-and-politics>; Alice Marwick, Rachel Kuo, Shanice Jones Cameron, and Moira Weigel, “Critical Disinformation Studies: A Syllabus,” Center for Information, Technology, and Public Life, 2021, <https://citap.unc.edu/research/critical-disinfo>; Ben Smith, “Inside the ‘Misinformation’ Wars,” *New York Times*, November 28, 2021, <https://www.nytimes.com/2021/11/28/business/media-misinformation-disinformation.html>; Matthew Yglesias, “The Misinformation Cope,” *Slow Boring*, April 20, 2022, <https://www.slowboring.com/p/misinformation>; Théophile Lenoir, “Reconsidering the Fight Against Disinformation,” *Tech Policy Press*, August 1, 2022, <https://www.techpolicy.press/reconsidering-the-fight-against-disinformation>; Dan Williams, “Misinformation Researchers Are Wrong: There Can’t Be a Science of Misleading Content,” *Conspicuous Cognition*, January 10, 2024, <https://www.conspicuouscognition.com/p/misinformation-researchers-are-wrong>; and Gavin Wilde, “From Panic to Policy: The Limits of Propaganda and the Foundations of an Effective Response,” *Texas National Security Review* (forthcoming 2024).
- 8 Jon Bateman, Elonnai Hickok, Laura Courchesne, Isra Thange, and Jacob N. Shapiro, “Measuring the Effects of Influence Operations: Key Findings and Gaps From Empirical

- Research,” June 28, 2021, Carnegie Endowment for International Peace, <https://carnegieendowment.org/2021/06/28/measuring-effects-of-influence-operations-key-findings-and-gaps-from-empirical-research-pub-84824>.
- 9 Consider David Salvo, Jamie Fly, and Laura Rosenberger, “The ASD Policy Blueprint for Countering Authoritarian Interference in Democracies,” German Marshall Fund, June 26, 2018, <https://www.gmfus.org/news/asd-policy-blueprint-countering-authoritarian-interference-democracies>; “A Multi-dimensional Approach to Disinformation,” European Commission, 2018, <https://op.europa.eu/en/publication-detail/-/publication/6ef4df8b-4cea-11e8-beld-01aa75ed71a1>; and Edward Lucas and Peter Pomeranzen, “Winning the Information War: Techniques and Counter-strategies to Russian Propaganda in Central and Eastern Europe,” Center for European Policy Analysis, 2016, https://cepa.ecms.pl/files/?id_plik=2773.
 - 10 Tom Stites, “A Quarter of All U.S. Newspapers Have Died in 15 Years, a New UNC News Deserts Study Found,” Poynter Institute, June 24, 2020, <https://www.poynter.org/locally/2020/unc-news-deserts-report-2020/>.
 - 11 Penelope Muse Abernathy, “News Deserts and Ghost Newspapers: Will Local News Survive?,” University of North Carolina, 2020, <https://www.usnewsdeserts.com/reports/news-deserts-and-ghost-newspapers-will-local-news-survive/>; and “The Tow Center COVID-19 Newsroom Cutback Tracker,” Tow Center for Digital Journalism, September 9, 2020, <https://www.cjr.org/widescreen/covid-cutback-tracker.php>.
 - 12 For example, see Dapo Olorunyomi, “Surviving the Pandemic: The Struggle for Media Sustainability in Africa,” National Endowment for Democracy, January 2021, <https://www.ned.org/wp-content/uploads/2021/01/Pandemic-Struggle-Media-Sustainability-Africa-Olorunyomi.pdf>.
 - 13 “About the Consortium,” New Jersey Civic Information Consortium, accessed January 27, 2023, <https://njcivicinfo.org/about/>.
 - 14 Anya Schiffrin, ed., *In the Service of Power: Media Capture and the Threat to Democracy*, Center for International Media Assistance, 2017, <https://www.cima.ned.org/resource/service-power-media-capture-threat-democracy/>.
 - 15 Consider “INN Mission & History,” Institute for Nonprofit News, accessed January 27, 2023, <https://inn.org/about/who-we-are/>.
 - 16 Jim Waterson, “VAT Ruling on Times Digital Edition Could Save News UK Millions,” *Guardian*, January 6, 2020, <https://www.theguardian.com/media/2020/jan/06/vat-ruling-on-times-digital-edition-could-save-news-uk-millions>.
 - 17 “About the Digital News Subscription Tax Credit,” Government of Canada, accessed March 24, 2023, <https://www.canada.ca/en/revenue-agency/services/tax/individuals/topics/about-your-tax-return/tax-return/completing-a-tax-return/deductions-credits-expenses/deductions-credits-expenses/digital-news-subscription.html>.
 - 18 “News Media Bargaining Code,” Australian Competition & Consumer Commission, accessed January 27, 2023, <https://www.accc.gov.au/by-industry/digital-platforms-and-services/news-media-bargaining-code/news-media-bargaining-code>.
 - 19 Julia Angwin, “Can Taxing Big Tech Save Journalism?” *Markup*, July 16, 2022, <https://themarkup.org/newsletter/hello-world/can-taxing-big-tech-save-journalism>.
 - 20 “INN Index 2022: Enduring in Crisis, Surging in Local Communities,” Institute for Nonprofit News, July 27, 2022, <https://inn.org/research/inn-index/inn-index-2022/>; and “Newsmatch,” Institute for Nonprofit News, accessed April 18, 2023, <https://newsmatch.inn.org/>.
 - 21 “INN Index 2022,” Institute for Nonprofit News.

- 22 “Newsmatch,” Institute for Nonprofit News.
- 23 “About Us,” Report for America, accessed January 27, 2023, <https://www.reportforamerica.org/about-us/>.
- 24 “Supporting Report for America,” Report for America, accessed December 23, 2023, <https://www.reportforamerica.org/supporters>.
- 25 Danny Hayes and Jennifer L. Lawless, “The Decline of Local News and Its Effects: New Evidence from Longitudinal Data,” *Journal of Politics* 80, no. 1 (January 2018): <https://www.dannyhayes.org/uploads/6/9/8/5/69858539/decline.pdf>.
- 26 The relationship between disinformation and trust in media, government, and other institutions is complex. Exposure to false content online is associated with lower trust in media but higher trust in government for conservatives when their preferred party is in power. Lack of trust in institutions is associated with higher belief in conspiracy theories, for example in the context of COVID-19 vaccination. See Katherine Ognyanova, David Lazer, Ronald E. Robertson, and Christo Wilson, “Misinformation in Action: Fake News Exposure Is Linked to Lower Trust in Media, Higher Trust in Government When Your Side Is in Power,” *Harvard Kennedy School Misinformation Review*, June 2, 2020, <https://misinforeview.hks.harvard.edu/article/misinformation-in-action-fake-news-exposure-is-linked-to-lower-trust-in-media-higher-trust-in-government-when-your-side-is-in-power>; and Will Jennings et al., “Lack of Trust, Conspiracy Beliefs, and Social Media Use Predict COVID-19 Vaccine Hesitancy,” *Vaccines* 9, no. 6 (June 2021): <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8226842>. See also Jay Jennings and Meghan Rubado, “Newspaper Decline and the Effect on Local Government Coverage,” University of Texas at Austin, November 2019, https://moody.utexas.edu/sites/default/files/Strauss_Research_Newspaper_Decline_2019-11-Jennings.pdf; Jackie Filla and Martin Johnson, “Local News Outlets and Political Participation,” *Urban Affairs Review* 45, no. 5 (2010): <https://journals.sagepub.com/doi/abs/10.1177/1078087409351947?journalCode=uarb>; “2021 Edelman Trust Barometer,” Edelman, 2021, <https://www.edelman.com/trust/2021-trust-barometer>; and Jeffrey Hiday, “Combating Disinformation by Bolstering Truth and Trust,” RAND Corporation, May 24, 2020, <https://www.rand.org/pubs/articles/2022/combating-disinformation-by-bolstering-truth-and-trust.html>.
- 27 Martin Baekgaard, Carsten Jensen, Peter B. Mortensen, and Søren Serritzlew, “Local News Media and Voter Turnout,” *Local Government Studies* 40 (2014): <https://www.tandfonline.com/doi/abs/10.1080/03003930.2013.834253>.
- 28 Christopher Chapp and Peter Aehl, “Newspapers and Political Participation: The Relationship Between Ballot Rolloff and Local Newspaper Circulation,” *Newspaper Research Journal* 42, no. 2 (2021): <https://journals.sagepub.com/doi/10.1177/07395329211014968>; and “Does Local Journalism Stimulate Voter Participation in State Supreme Court Elections?,” David Hughes, *Journal of Law and Courts* 8, no. 1 (2020): <https://www.cambridge.org/core/journals/journal-of-law-and-courts/article/abs/does-local-journalism-stimulate-voter-participation-in-state-supreme-court-elections/CE8E2CBDF4CF9C58DF08A013AE8B05A3>.
- 29 Matthew Gentzkow, Jesse M. Shapiro, and Michael Sinkinson, “The Effect of Newspaper Entry and Exit on Electoral Politics,” *American Economic Review* 101 (December 2011): <https://web.stanford.edu/~gentzkow/research/voting.pdf>. For a roundup of this research, see Josh Stearns and Christine Schmidt, “How We Know Journalism Is Good for Democracy,” Democracy Fund, September 15, 2022, <https://democracyfund.org/idea/how-we-know-journalism-is-good-for-democracy/>.

- 30 David S. Ardia, Evan Ringel, Victoria Ekstrand, and Ashley Fox, “Addressing the Decline of Local News, Rise of Platforms, and Spread of Mis- and Disinformation Online: A Summary of Current Research and Policy Proposals,” University of North Carolina, December 22, 2020, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3765576.
- 31 Jessica Mahone and Philip Napoli, “Hundreds of Hyperpartisan Sites Are Masquerading as Local News. This Map Shows If There’s One Near You,” Nieman Lab, July 13, 2020, <https://www.niemanlab.org/2020/07/hundreds-of-hyperpartisan-sites-are-masquerading-as-local-news-this-map-shows-if-theres-one-near-you/>.
- 32 Joshua P. Darr, Matthew P. Hitt, and Johanna L. Dunaway, “Newspaper Closures Polarize Voting Behavior,” *Journal of Communication* 68, no. 6 (December 2018): <https://academic.oup.com/joc/article-abstract/68/6/1007/5160090>.
- 33 See Imelda Deinla, Gabrielle Ann S. Mendoza, Kier Jesse Ballar, and Jurel Yap, “The Link Between Fake News Susceptibility and Political Polarization of the Youth in the Philippines,” Ateneo School of Government, Working Paper no. 21-029, November 2021, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3964492; and Mathias Osmundsen, Michael Bang Petersen, and Alexander Bor, “How Partisan Polarization Drives the Spread of Fake News,” Brookings Institution, May 13, 2021, <https://www.brookings.edu/articles/how-partisan-polarization-drives-the-spread-of-fake-news/>. In the United States, Yochai Benkler and others have argued that asymmetric polarization—with the right drifting from the center faster and farther than the left—has been driven at least in part by media dynamics. Specifically, right-leaning media across cable television, radio, and the internet are less connected to mainstream media than their left-leaning counterparts. See Yochai Benkler, Robert Faris, Hal Roberts, and Ethan Zuckerman, “Study: Breitbart-Led Right-Wing Media Ecosystem Altered Broader Media Agenda,” *Columbia Journalism Review*, March 3, 2017, <https://www.cjr.org/analysis/breitbart-media-trump-harvard-study.php>. Not all analysts believe polarization is a bad thing; for example, some argue that polarization provides voters more distinct choices and has led to increased political participation. Others have argued that polarization contributes to crisis flash points that disrupt a problematic status quo in ways that are ultimately healthy. Consider Jessica Rettig, “Why Political Polarization Might Be Good for America,” *U.S. News*, May 27, 2010, <https://www.usnews.com/opinion/articles/2010/05/27/why-political-polarization-might-be-good-for-america>; see also “The US Is Suffering From Toxic Polarization. That’s Arguably a Good Thing,” Peter T. Coleman, *Scientific American*, April 2, 2021, <https://www.scientificamerican.com/article/the-u-s-is-suffering-from-toxic-polarization-thats-arguably-a-good-thing>. The United States is an international outlier on polarization. A review by Jennifer McCoy and Benjamin Press found that the United States is “the only advanced Western democracy to have faced such intense polarization for such an extended period.” Their study suggests grim outcomes from high levels of polarization. McCoy and Press examine a sample of fifty-two democratic societies suffering “pernicious polarization,” defined “as the division of society into mutually distrustful political camps in which political identity becomes a social identity.” They find that half of cases faced democratic erosion, and fewer than a fifth were able to sustain a decline in pernicious polarization. Jennifer McCoy and Benjamin Press, “What Happens When Democracies Become Perniciously Polarized?” *Carnegie Endowment for International Peace*, January 18, 2022, <https://carnegieendowment.org/2022/01/18/what-happens-when-democracies-become-perniciously-polarized-pub-86190>.
- 34 Stites, “Quarter of All U.S. Newspapers”; “Fiscal Year 2022 Operating Budget,” Corporation for Public Broadcasting, accessed January 27, 2023, <https://www.cpb.org/aboutcpb/financials/budget>; Anthony P. Carnevale and Emma Wenzinger, “Stop the Presses: Journalism

- Employment and Economic Value of 850 Journalism and Communication Programs,” Georgetown University Center on Education and Workforce, 2022, <https://cewgeorgetown.wpenginepowered.com/wp-content/uploads/cew-journalism-fr.pdf>.
- 35 Perry Bacon, Jr., “America Should Spend Billions to Revive Local News,” *Washington Post*, October 17, 2022, <https://www.washingtonpost.com/opinions/2022/10/17/local-news-crisis-plan-fix-perry-bacon/>.
 - 36 “National Ecosystem Calendar,” Democracy Fund, accessed April 18, 2023, https://oneroyalace.github.io/news-ecosystem-model/national_calculator.html.
 - 37 See Brian Fung, “Meta Avoids Showdown Over News Content in US After Journalism Bargaining Bill Shelved,” CNN, December 7, 2022, <https://www.cnn.com/2022/12/07/tech/meta-journalism-bargaining-bill/index.html>; Joshua Benton, “Don’t Expect McConnell’s Paradox to Help News Publishers Get Real Money Out of Google and Facebook,” Nieman Lab, January 8, 2020, <https://www.niemanlab.org/2020/01/dont-expect-mcconnells-paradox-to-help-news-publishers-get-real-money-out-of-google-and-facebook/>; Jeff Jarvis, “As Rupert Murdoch Works to Dismantle the Internet, Why Are Other Media Outlets Helping Him?,” *Crikey*, February 15, 2021, <https://www.crikey.com.au/2021/02/15/rupert-murdoch-news-media-bargaining-code/>; Josh Frydenberg, “Review of the News Media and Digital Platforms Mandatory Bargaining Code,” Australian Department of the Treasury, February 2022, <https://ministers.treasury.gov.au/ministers/josh-frydenberg-2018/media-releases/review-news-media-and-digital-platforms-mandatory>; and Anya Schiffrin, “Australia’s News Media Bargaining Code Pries \$140 Million From Google and Facebook,” Poynter Institute, August 16, 2022, <https://www.poynter.org/business-work/2022/australias-news-media-bargaining-code-pries-140-million-from-google-and-facebook>.
 - 38 Max Matza, “Google and Canada Reach Deal to Avert News Ban Over Online News Act,” BBC, November 29, 2023, <https://www.bbc.com/news/world-us-canada-67571027>; and Jaimie Ding, “California Bill Requiring Big Tech to Pay for News Placed on Hold Until 2024,” *Los Angeles Times*, July 7, 2023, <https://www.latimes.com/business/story/2023-07-07/california-journalism-bill-on-hold-until-2024>.
 - 39 For examples, see Lucas and Pomeranzev, “Winning the Information War”; Katarína Klingová and Daniel Milo, “Countering Information War Lessons Learned from NATO and Partner Countries: Recommendations and Conclusions,” GLOBSEC, February 2017, <https://www.globsec.org/what-we-do/publications/countering-information-war-lessons-learned-nato-and-partner-countries>; Claire Wardle and Hossein Derakhshan, “Information Disorder: Toward an Interdisciplinary Framework for Research and Policy Making,” Council of Europe, September 2017, <https://rm.coe.int/information-disorder-toward-an-interdisciplinary-framework-for-research/168076277c>; Daniel Fried and Alina Polyakova, “Democratic Defense Against Disinformation,” Atlantic Council, February 2018, <https://www.atlanticcouncil.org/wp-content/uploads/2018/03/Democratic-Defense-Against-Disinformation-FINAL.pdf>; “A Multi-Dimensional Approach,” European Commission; Erik Brattberg and Tim Maurer, “Russian Election Interference: Europe’s Counter to Fake News and Cyber Attacks,” Carnegie Endowment for International Peace, May 2018, https://carnegieendowment.org/files/CP_333_BrattbergMaurer_Russia_Elections_Interference_FINAL.pdf; “Action Plan Against Disinformation,” European Commission, May 2018, https://www.eecas.europa.eu/node/54866_en; Fly, Rosenberger, and Salvo, “The ASD Policy Blueprint”; Jean-Baptiste Jeangène Vilmer, Alexandre Escorcía, Marine Guillaume, and Janaina Herrera, “Information Manipulation: A Challenge for Our Democracies,” French Ministry for Europe and Foreign Affairs and the Institute for Strategic Research, August 2018, https://www.diplomatie.gouv.fr/IMG/pdf/information_manipulation_rvb_cle838736.pdf; Todd C. Helmus et al., “Russian

- Social Media Influence: Understanding Russian Propaganda in Eastern Europe,” RAND Corporation, 2018, https://www.rand.org/pubs/research_reports/RR2237.html; and Paul Barrett, “Tackling Domestic Disinformation: What the Social Media Companies Need to Do,” New York University, March 2019, https://issuu.com/nyusterncenterforbusinessandhumanri/docs/nyu_domestic_disinformation_digital?e=31640827/68184927.
- 40 Klingová and Milo, “Countering Information War.”
 - 41 “Media Literacy Defined,” National Association for Media Literacy Education, accessed February 13, 2023, <https://namle.net/resources/media-literacy-defined/>. See also Monica Bulger and Patrick Davison, “The Promises, Challenges, and Futures of Media Literacy,” Data & Society, February 21, 2018, <https://datasociety.net/library/the-promises-challenges-and-futures-of-media-literacy/>; and Géraldine Wuyckens, Normand Landry, and Pierre Fastrez, “Untangling Media Literacy, Information Literacy, and Digital Literacy: A Systematic Meta-review of Core Concepts in Media Education,” *Journal of Media Literacy Education* 14, no. 1 (2022): <https://digitalcommons.uri.edu/cgi/viewcontent.cgi?article=1531&context=jmle>.
 - 42 Renee Hobbs, “Digital and Media Literacy: A Plan of Action,” Aspen Institute, 2010, https://www.aspeninstitute.org/wp-content/uploads/2010/11/Digital_and_Media_Literacy.pdf; and “Online Media Literacy Strategy,” UK Department for Digital, Culture, Media, & Sport, July 2021, https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/1004233/DCMS_Media_Literacy_Report_Roll_Out_Accessible_PDF.pdf.
 - 43 Tiffany Hsu, “When Teens Find Misinformation, These Teachers Are Ready,” *New York Times*, September 8, 2022, <https://www.nytimes.com/2022/09/08/technology/misinformation-students-media-literacy.html>; “A Global Study on Information Literacy: Understanding Generational Behaviors and Concerns Around False and Misleading Information Online,” Poynter Institute, August 2022, <https://www.poynter.org/wp-content/uploads/2022/08/A-Global-Study-on-Information-Literacy-1.pdf>; and Elena-Alexandra Dumitru, “Testing Children and Adolescents’ Ability to Identify Fake News: A Combined Design of Quasi-Experiment and Group Discussions,” *Societies* 10, no. 3 (September 2020): <https://www.mdpi.com/2075-4698/10/3/71/html>.
 - 44 Sam Wineburg, Sarah McGrew, Joel Breakstone, and Teresa Ortega, “Evaluating Information: The Cornerstone of Civic Online Reasoning,” Stanford Digital Repository, November 22, 2016, <https://purl.stanford.edu/fv751yt5934>.
 - 45 For evidence that older users are more likely to share false stories on Facebook, see Andrew Guess, Jonathan Nagler, and Joshua Tucker, “Less Than You Think: Prevalence and Predictors of Fake News Dissemination on Facebook,” *Science Advances* 5, no. 1 (2019): <https://www.science.org/doi/10.1126/sciadv.aau4586>.
 - 46 Elisabeth Braw, “Create a Psychological Defence Agency to ‘Prebunk’ Fake News,” *Prospect*, December 8, 2022, <https://www.prospectmagazine.co.uk/politics/60291/create-a-psychological-defence-agency-to-prebunk-fake-news>; and Adela Suliman, “Sweden Sets Up Psychological Defense Agency to Fight Fake News, Foreign Interference,” *Washington Post*, January 6, 2022, <https://www.washingtonpost.com/world/2022/01/06/sweden-fake-news-psychological-defence-agency>.
 - 47 Erin Murrock, Joy Amulya, Mehri Druckman, and Tetiana Liubyva, “Winning the War on State-Sponsored Propaganda: Gains in the Ability to Detect Disinformation a Year and a Half After Completing a Ukrainian News Media Literacy Program,” *Journal of Media Literacy Education* 10, no. 2 (2018): <https://digitalcommons.uri.edu/cgi/viewcontent.cgi?article=1361&context=jmle>.

- 48 “Online Media Literacy Strategy,” UK Department for Digital, Culture, Media, & Sport; and Kara Brisson-Boivin and Samantha McAleese, “From Access to Engagement: Building a Digital Media Literacy Strategy for Canada,” MediaSmarts, 2022, <https://mediasmarts.ca/research-reports/access-engagement-building-digital-media-literacy-strategy-canada>.
- 49 Hasan Tinmaz, Yoo-Taek Lee, Mina Fanea-Ivanovici, and Hasnan Baber, “A Systematic Review on Digital Literacy,” *Smart Learning Environments* 9 (2022), <https://slejournal.springeropen.com/articles/10.1186/s40561-022-00204-y>.
- 50 “History,” National Association for Media Literacy Education, accessed February 13, 2023, <https://namle.net/about/history/>.
- 51 “Media & Information Literacy,” UN Alliance of Civilizations, accessed March 26, 2023, <https://milunesco.unaoc.org/mil-organizations/acma-digital-media-literacy-research-program>.
- 52 “Media & Information Literacy,” UN Alliance of Civilizations.
- 53 “Online Media Literacy Strategy,” UK Department for Digital, Culture, Media, & Sport; “Media Literacy Programme Fund,” Government of the United Kingdom, accessed March 26, 2023, <https://www.gov.uk/guidance/media-literacy-programme-fund>; and Bulger and Davison, “Promises, Challenges, and Futures.”
- 54 Consider Bulger and Davison, “Promises, Challenges, and Futures,” as well as Theodora Dame Adjin-Tettey, “Combating Fake News, Disinformation, and Misinformation: Experimental Evidence for Media Literacy Education,” *Cogent Arts & Humanities* 9 (2022): <https://www.randfonline.com/doi/full/10.1080/23311983.2022.2037229>.
- 55 Jon Roozenbeek and Sander van der Linden, “Fake News Game Confers Psychological Resistance Against Online Misinformation,” *Palgrave Communications* 5 (2019): <https://www.nature.com/articles/s41599-019-0279-9>.
- 56 Murrock, Amulya, Druckman, and Liubyva, “Winning the War.”
- 57 Carl-Anton Werner Axelsson, Mona Guath, and Thomas Nygren, “Learning How to Separate Fake From Real News: Scalable Digital Tutorials Promoting Students’ Civic Online Reasoning,” *Future Internet* 13, no. 3 (2021): <https://www.mdpi.com/1999-5903/13/3/60>.
- 58 Jennifer Fleming, “Media Literacy, News Literacy, or News Appreciation? A Case Study of the News Literacy Program at Stony Brook University,” *Journalism & Mass Communication Educator* 69, no. 2 (2013): <https://journals.sagepub.com/doi/abs/10.1177/1077695813517885>.
- 59 Because the measurement of media literacy was self-reported, the study posits this as an example of the “Dunning-Kruger effect”: an individual’s (over)confidence in their ability to critically consume media is related to their susceptibility to deception. See Mo Jones-Jang, Tara Mortensen, and Jingjing Liu, “Does Media Literacy Help Identification of Fake News? Information Literacy Helps, but Other Literacies Don’t,” *American Behavioral Scientist* (August 2019): https://www.researchgate.net/publication/335352499_Does_Media_Literacy_Help_Identification_of_Fake_News_Information_Literacy_Helps_but_Other_Literacies_Don't.
- 60 Brigitte Huber, Porismita Borah, and Homero Gil de Zúñiga, “Taking Corrective Action When Exposed to Fake News: The Role of Fake News Literacy,” *Journal of Media Literacy Education* 14 (July 2022): https://www.researchgate.net/publication/362513295_Taking_corrective_action_when_exposed_to_fake_news_The_role_of_fake_news_literacy.
- 61 Murrock, Amulya, Druckman, and Liubyva, “Winning the War”; and “Impact Report: Evaluating PEN America’s Media Literacy Program,” PEN America & Stanford Social Media Lab, September 2022, <https://pen.org/report/the-impact-of-community-based-digital-literacy-interventions-on-disinformation-resilience>. See also Yan Su, Danielle Ka Lai Lee, and Xizhu Xiao, “‘I Enjoy Thinking Critically, and I’m in Control’: Examining the Influences of Media

- Literacy Factors on Misperceptions Amidst the COVID-19 Infodemic,” *Computers in Human Behavior* 128 (2022): <https://www.sciencedirect.com/science/article/pii/S0747563221004349>, a study based on subjects in China. The similar findings between the United States, Ukraine, and China—despite significant differences in the three countries’ media systems and histories—is noteworthy.
- 62 See generally: Folco Panizza et al., “Lateral Reading and Monetary Incentives to Spot Disinformation About Science,” *Scientific Reports* 12 (2022): <https://www.nature.com/articles/s41598-022-09168-y>; Sam Wineburg et al., “Lateral Reading on the Open Internet: A District-Wide Field Study in High School Government Classes,” *Journal of Educational Psychology* 114, no. 5 (2022): <https://www.studocu.com/id/document/universitas-kristen-satya-wacana/social-psychology/lateral-reading-on-the-open-internet-a-district-wide-field-study-in-high-school-government-classes/45457099>; and Joel Breakstone et al., “Lateral Reading: College Students Learn to Critically Evaluate Internet Sources in an Online Course,” *Harvard Kennedy School Misinformation Review* 2 (2021), <https://misinforeview.hks.harvard.edu/article/lateral-reading-college-students-learn-to-critically-evaluate-internet-sources-in-an-online-course>.
 - 63 For more on this method, its success in classroom trials, and its departure from previous forms of media literacy education, see D. Pavlounis, J. Johnston, J. Brodsky, and P. Brooks, “The Digital Media Literacy Gap: How to Build Widespread Resilience to False and Misleading Information Using Evidence-Based Classroom Tools,” CIVIX Canada, November 2021, <https://ctrl-f.ca/en/wp-content/uploads/2021/11/The-Digital-Media-Literacy-Gap-Nov-7.pdf>.
 - 64 Axelsson, Guath, and Nygren, “Learning How to Separate.”
 - 65 Jessica E. Brodsky et al., “Associations Between Online Instruction in Lateral Reading Strategies and Fact-Checking COVID-19 News Among College Students,” *AERA Open* (2021): <https://journals.sagepub.com/doi/full/10.1177/23328584211038937>.
 - 66 Sarah McGrew, Mark Smith, Joel Breakstone, Teresa Ortega, and Sam Wineburg, “Improving University Students’ Web Savvy: An Intervention Study,” *British Journal of Educational Psychology* 89, no. 3 (September 2019): <https://bpspsychub.onlinelibrary.wiley.com/doi/10.1111/bjep.12279>.
 - 67 Angela Kohnen, Gillian Mertens, and Shelby Boehm, “Can Middle Schoolers Learn to Read the Web Like Experts? Possibilities and Limits of a Strategy-Based Intervention,” *Journal of Media Literacy Education* 12, no. 2 (2020): <https://digitalcommons.uri.edu/cgi/viewcontent.cgi?article=1457&context=jmle>.
 - 68 Adam Maksl, Seth Ashley, and Stephanie Craft, “Measuring News Media Literacy,” *Journal of Media Literacy Education* 6 (2015), <https://digitalcommons.uri.edu/jmle/vol6/iss3/3/>; and Huber, Borah, and Zúñiga, “Taking Corrective Action.”
 - 69 Bulger and Davison, “Promises, Challenges, and Futures.”
 - 70 Like Jones-Jang, Mortensen, and Liu, the authors of the IREX evaluation suggest that the “false sense of control” already felt by individuals who did not receive media literacy training may also partially explain the relatively small improvements in these subjects’ locus of control.
 - 71 Paul Mihailidis, “Beyond Cynicism: Media Education and Civic Learning Outcomes in the University,” *International Journal of Learning and Media* 1, no. 3 (August 2009): https://www.researchgate.net/publication/250958225_Beyond_Cynicism_Media_Education_and_Civic_Learning_Outcomes_in_the_University.
 - 72 “You Think You Want Media Literacy... Do You?” danah boyd, apophenia, March 9, 2018, <https://www.zephoria.org/thoughts/archives/2018/03/09/you-think-you-want-media-literacy-do-you.html>.

- 73 Hobbs, “Digital and Media Literacy.”
- 74 Sandy Zinn, Christine Stilwell, and Ruth Hoskins, “Information Literacy Education in the South African Classroom: Reflections from Teachers’ Journals in the Western Cape Province,” *Libri* 66 (April 2016): <https://www.degruyter.com/document/doi/10.1515/libri-2015-0102/html>; and Maria Ranieri, Isabella Bruni, and Anne-Claire Orban de Xivry, “Teachers’ Professional Development on Digital and Media Literacy. Findings and Recommendations From a European Project,” *Research on Education and Media* 9, no. 2 (2017): <https://sciendo.com/article/10.1515/rem-2017-0009>.
- 75 Victoria Smith, “Mapping Worldwide Initiatives to Counter Influence Operations,” Carnegie Endowment for International Peace, December 14, 2020, <https://carnegieendowment.org/2020/12/14/mapping-worldwide-initiatives-to-counter-influence-operations-pub-83435>; and “Fact-Checking,” Duke Reporter’s Lab, accessed January 27, 2023, <https://reporterslab.org/fact-checking>.
- 76 “The CoronaVirusFacts/DatosCoronaVirus Alliance Database,” Poynter Institute, accessed December 10, 2023, <https://www.poynter.org/ifcn-covid-19-misinformation>.
- 77 “Verificado 2018,” Online Journalism Awards, accessed December 10, 2023, <https://awards.journalists.org/entries/verificado-2018>.
- 78 “About Fact-Checking on Facebook and Instagram,” Meta, accessed March 22, 2023, <https://www.facebook.com/business/help/2593586717571940?id=673052479947730>.
- 79 John M. Carey et al., “The Effects of Corrective Information About Disease Epidemics and Outbreaks: Evidence From Zika and Yellow Fever in Brazil,” *Science Advances* 6 (2020), <https://www.science.org/doi/10.1126/sciadv.aaw7449>; Jeremy Bowles, Horacio Larreguy, and Shelley Liu, “Countering Misinformation via WhatsApp: Preliminary Evidence From the COVID-19 Pandemic in Zimbabwe,” *PLOS ONE* 15 (2020), <https://doi.org/10.1371/journal.pone.0240005>; and Sara Pluviano, Sergio Della Sala, and Caroline Watt, “The Effects of Source Expertise and Trustworthiness on Recollection: The Case of Vaccine Misinformation,” *Cognitive Processing* 21 (2020), <https://pubmed.ncbi.nlm.nih.gov/32333126/>.
- 80 See Brendan Nyhan, Ethan Porter, Jason Reifler, and Thomas Wood, “Taking Fact-Checks Literally but Not Seriously? The Effects of Journalistic Fact-Checking on Factual Beliefs and Candidate Favorability,” *Political Behavior* 42 (2019): <https://link.springer.com/article/10.1007/s11109-019-09528-x>; Briony Swire-Thompson, Ullrich K. H. Ecker, Stephan Lewandowsky, and Adam J. Berinsky, “They Might Be a Liar But They’re My Liar: Source Evaluation and the Prevalence of Misinformation,” *Political Psychology* 41 (2020), <https://onlinelibrary.wiley.com/doi/abs/10.1111/pops.12586>; Oscar Barrera, Sergei Guriev, Emeric Henry, and Ekaterina Zhuravskaya, “Facts, Alternative Facts, and Fact Checking in Times of Post-Truth Politics,” *Journal of Public Economics* 182 (2017), <https://www.sciencedirect.com/science/article/pii/S0047272719301859>; and Briony Swire, Adam J. Berinsky, Stephan Lewandowsky, and Ullrich K. H. Ecker, “Processing Political Misinformation: Comprehending the Trump Phenomenon,” *Royal Society Open Science* 4 (2017), <https://royalsocietypublishing.org/doi/10.1098/rsos.160802>.
- 81 Antino Kim and Alan R. Dennis, “Says Who? The Effects of Presentation Format and Source Rating on Fake News in Social Media,” *MIS Quarterly* 43, no. 3 (2019): https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2987866; Ethan Porter, Thomas J. Wood, and David Kirby, “Sex Trafficking, Russian Infiltration, Birth Certificates, and Pedophilia: A Survey Experiment Correcting Fake News,” *Journal of Experimental Political Science* 5, no. 2 (2018): <https://www.cambridge.org/core/journals/journal-of-experimental-political-science/article/sex-trafficking-russian-infiltration-birth-certificates-and-pedophilia-a-survey-experiment-correcting-fake-news/CFEB9AFD5F0AEB64DF32D5A7641805B6>; and Jeong-woo Jang, Eun-Ju Lee, and Soo Yun

- Shin, "What Debunking of Misinformation Does and Doesn't," *Cyberpsychology, Behavior, and Social Networking* 22, no. 6 (2019): <https://pubmed.ncbi.nlm.nih.gov/31135182>.
- 82 Jimmeka J. Guillory and Lisa Geraci "Correcting Erroneous Inferences in Memory: The Role of Source Credibility," *Journal of Applied Research in Memory and Cognition* 2, no. 4 (2013): <https://doi.org/10.1016/j.jarmac.2013.10.001>; and Pluviano and Della Sala, "Effects of Source Expertise."
 - 83 Maurice Jakesch, Moran Koren, Anna Evtushenko, and Mor Naaman, "The Role of Source, Headline and Expressive Responding in Political News Evaluation," SSRN, January 31, 2019, <https://dx.doi.org/10.2139/ssrn.3306403>.
 - 84 Consider Thomas Carothers and Andrew O'Donohue, "How Americans Were Driven to Extremes: In the United States, Polarization Runs Particularly Deep," *Foreign Affairs*, September 25, 2019, <https://www.foreignaffairs.com/articles/united-states/2019-09-25/how-americans-were-driven-extremes>.
 - 85 Consider Michael J. Aird, Ullrich K. H. Ecker, Briony Swire, Adam J. Berinsky, and Stephan Lewandowsky, "Does Truth Matter to Voters? The Effects of Correcting Political Misinformation in an Australian Sample," *Royal Society Open Science* (2018), <https://royalsocietypublishing.org/doi/10.1098/rsos.180593>.
 - 86 The backfire effect was captured in Brendan Nyhan and Jason Reifler, "When Corrections Fail: The Persistence of Political Misperceptions," *Political Behavior* 32 (2010): <https://link.springer.com/article/10.1007/s11109-010-9112-2>. The popular online comic *The Oatmeal* featured commentary about the backfire effect, demonstrating its breakthrough into popular imagination; see "Believe," *The Oatmeal*, accessed January 27, 2023, <https://theoatmeal.com/comics/believe>. However, other studies have since called the effect into question. See Thomas Wood and Ethan Porter, "The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence," *Political Behavior* 41 (2019): <https://link.springer.com/article/10.1007/s11109-018-9443-y>; see also Kathryn Haglin, "The Limitations of the Backfire Effect," *Research & Politics* 4 (2017): <https://journals.sagepub.com/doi/10.1177/2053168017716547>; and Brendan Nyhan, "Why the Backfire Effect Does Not Explain the Durability of Political Misperceptions," *PNAS* 118 (2020): <https://www.pnas.org/doi/10.1073/pnas.1912440117>.
 - 87 Brendan Nyhan and Jason Reifler, "Displacing Misinformation About Events: An Experimental Test of Causal Corrections," *Journal of Experimental Political Science* 2 (2015): <https://www.cambridge.org/core/journals/journal-of-experimental-political-science/article/abs/displacing-misinformation-about-events-an-experimental-test-of-causal-corrections/69550AB61F4E3F7C2CD03532FC740D05>.
 - 88 Angeline Sangalang, Yotam Ophir, and Joseph N. Cappella, "The Potential for Narrative Correctives to Combat Misinformation," *Journal of Communication* 69, no. 3 (2019): <https://academic.oup.com/joc/article-abstract/69/3/298/5481803?redirectedFrom=fulltext>.
 - 89 Brian E. Weeks, "Emotions, Partisanship, and Misperceptions: How Anger and Anxiety Moderate the Effect of Partisan Bias on Susceptibility to Political Misinformation," *Journal of Communication* 65, no. 4 (2015): <https://onlinelibrary.wiley.com/doi/abs/10.1111/jcom.12164>.
 - 90 Ethan Porter and Thomas Wood, "The Global Effectiveness of Fact-Checking: Evidence From Simultaneous Experiments in Argentina, Nigeria, South Africa, and the United Kingdom," *PNAS* 118, no. 37 (2021): <https://www.pnas.org/doi/10.1073/pnas.2104235118>; see also John M. Carey et al., "The Ephemeral Effects of Fact-Checks on COVID-19 Misperceptions in the United States, Great Britain and Canada," *Nature Human Behaviour* 6 (2022), <https://www.nature.com/articles/s41562-021-01278-3>; and Patrick R. Rich and Maria S. Zaragoza, "Correcting Misinformation in News Stories: An Investigation of Correction Timing and

- Correction Durability,” *Journal of Applied Research in Memory and Cognition* 9, no. 3 (2020): <https://www.sciencedirect.com/science/article/abs/pii/S2211368120300280>.
- 91 Emily Thorson, “Belief Echoes: The Persistent Effects of Corrected Misinformation,” *Political Communication* 33, no. 3 (2015): <https://www.tandfonline.com/doi/full/10.1080/10584609.2015.1102187>.
 - 92 Consider Mevan Babakar, “Crowdsourced Factchecking: A Pie in The Sky?” *European Journalism Observatory*, June 1, 2018, <https://en.ejo.ch/specialist-journalism/crowdsourced-factchecking-a-pie-in-the-sky>. Studies suggest interventions from users can be as or more effective than interventions from experts: consider Leticia Bode and Emily K. Vraga, “See Something, Say Something: Correction of Global Health Misinformation on Social Media,” *Health Communication* 33, no. 9 (2018): <https://www.tandfonline.com/doi/full/10.1080/10410236.2017.1331312>; and Jonas Colliander, “This is Fake News: Investigating the Role of Conformity to Other Users’ Views When Commenting on and Spreading Disinformation in Social Media,” *Computers in Human Behavior* 97 (August 2019): <https://linkinghub.elsevier.com/retrieve/pii/S074756321930130X>.
 - 93 Sam Guzik, “AI Will Start Fact-Checking. We May Not Like the Results,” Nieman Lab, December 2022, <https://www.niemanlab.org/2022/12/ai-will-start-fact-checking-we-may-not-like-the-results>; and Grace Abels, “Can ChatGPT Fact-Check? We Tested,” Nieman Lab, May 31, 2023, <https://www.poynter.org/fact-checking/2023/chatgpt-ai-replace-fact-checking>.
 - 94 Mihai Avram, Nicholas Micallef, Sameer Patil, and Filippo Menczer, “Exposure to Social Engagement Metrics Increases Vulnerability to Misinformation,” *Harvard Kennedy School Misinformation Review* 1 (2020), <https://misinformreview.hks.harvard.edu/article/exposure-to-social-engagement-metrics-increases-vulnerability-to-misinformation>.
 - 95 Brian Stelter, “Facebook to Start Putting Warning Labels on ‘Fake News’,” CNN, December 15, 2016, <https://money.cnn.com/2016/12/15/media/facebook-fake-news-warning-labels>.
 - 96 For data on the rise of labeling and redirection, see Kamya Yadav, “Platform Interventions: How Social Media Counters Influence Operations,” Carnegie Endowment for International Peace, January 25, 2021, <https://carnegieendowment.org/2021/01/25/platform-interventions-how-social-media-counters-influence-operations-pub-83698>.
 - 97 Björn Ross, Anna-Katharina Jung, Jennifer Heisel, and Stefan Stieglitz, “Fake News on Social Media: The (In)Effectiveness of Warning Messages” (paper presented at Thirty-Ninth International Conference on Information Systems, San Francisco, 2018), https://www.researchgate.net/publication/328784235_Fake_News_on_Social_Media_The_InEffectiveness_of_Warning_Messages.
 - 98 “Policy Advisory Opinion 2022-01, Removal of COVID-19 Misinformation,” Oversight Board, April 2023, <https://oversightboard.com/attachment/547865527461223>.
 - 99 “Rating Process and Criteria,” NewsGuard, accessed February 7, 2023, <https://www.newsguardtech.com/ratings/rating-process-criteria>.
 - 100 Kevin Aslett et al., “News Credibility Labels Have Limited Average Effects on News Diet Quality and Fail to Reduce Misperceptions,” *Science Advances* 8, no. 18 (2022): <https://www.science.org/doi/10.1126/sciadv.abl3844>.
 - 101 Alexander Bor et al., “‘Fact-Checking’ Videos Reduce Belief in, but Not the Sharing of Fake News on Twitter,” PsyArXiv, April 11, 2020, <https://osf.io/preprints/psyarxiv/a7huq>.
 - 102 Gordon Pennycook et al., “Shifting Attention to Accuracy Can Reduce Misinformation Online,” *Nature* 592 (2021): <https://www.nature.com/articles/s41586-021-03344-2>; Gordon

- Pennycook et al., “Fighting COVID-19 Misinformation on Social Media: Experimental Evidence for a Scalable Accuracy-Nudge Intervention,” *Psychological Science* 31, no. 7 (2020): <https://journals.sagepub.com/doi/full/10.1177/0956797620939054>.
- 103 Timo K. Koch, Lena Frischlich, and Eva Lerner, “Effects of Fact-Checking Warning Labels and Social Endorsement Cues on Climate Change Fake News Credibility and Engagement on Social Media,” *Journal of Applied Social Psychology* 53, no. 3 (June 2023): <https://onlinelibrary.wiley.com/doi/10.1111/jasp.12959?af=R>; and Megan Duncan, “What’s in a Label? Negative Credibility Labels in Partisan News,” *Journalism & Mass Communication Quarterly* 99, no. 2 (2020): <https://journals.sagepub.com/doi/10.1177/1077699020961856?icid=int.sj-full-text-citing-articles.17>.
 - 104 Megan A. Brown et al., “Twitter Put Warning Labels on Hundreds of Thousands of Tweets. Our Research Examined Which Worked Best,” *Washington Post*, December 9, 2020, <https://www.washingtonpost.com/politics/2020/12/09/twitter-put-warning-labels-hundreds-thousands-tweets-our-research-examined-which-worked-best>.
 - 105 Gordon Pennycook, Adam Bear, Evan T. Collins, and David G. Rand, “The Implied Truth Effect: Attaching Warnings to a Subset of Fake News Headlines Increases Perceived Accuracy of Headlines Without Warnings,” *Management Science* 66, no. 11 (November 2020): <https://pubsonline.informs.org/doi/10.1287/mnsc.2019.3478>.
 - 106 Antino Kim, Patricia L. Moravec, and Alan R. Dennis, “Combating Fake News on Social Media With Source Ratings: The Effects of User and Expert Reputation Ratings,” *Journal of Management Information Systems* 36, no. 3 (2019): <https://www.tandfonline.com/doi/full/10.1080/07421222.2019.1628921>.
 - 107 Jan Kirchner and Christian Reuter, “Countering Fake News: A Comparison of Possible Solutions Regarding User Acceptance and Effectiveness,” *Proceedings of the ACM on Human-Computer Interaction* 4 (October 2020): https://www.peasec.de/paper/2020/2020_KirchnerReuter_CounteringFakeNews_CSCW.pdf; and Ciarra N. Smith and Holli H. Seitz, “Correcting Misinformation About Neuroscience via Social Media,” *Science Communication* 41, no. 6 (2019): <https://journals.sagepub.com/doi/10.1177/1075547019890073>.
 - 108 Avram, Micallef, Patil, and Menczer, “Exposure to Social Engagement Metrics.”
 - 109 Yadav, “Platform Interventions.”
 - 110 For an example of research that can be conducted with this data, see Samantha Bradshaw and Shelby Grossman, “Were Facebook and Twitter Consistent in Labeling Misleading Posts During the 2020 Election?” *Lawfare*, August 7, 2022, <https://www.lawfaremedia.org/article/were-facebook-and-twitter-consistent-labeling-misleading-posts-during-2020-election>.
 - 111 See Laura Livingston, “Understanding the Context Around Content: Looking Behind Misinformation Narratives,” National Endowment for Democracy, December 2021, <https://www.ned.org/wp-content/uploads/2021/12/Understanding-the-Context-Around-Content-Looking-Behind-Misinformation-Narratives-Laura-Livingston.pdf>; and Rachel Brown and Laura Livingston, “Counteracting Hate and Dangerous Speech Online: Strategies and Considerations,” Toda Peace Institute, March 2019, https://toda.org/assets/files/resources/policy-briefs/t-pb-34_brown-and-livingston_counteracting-hate-and-dangerous-speech-online.pdf. Additionally, consider Claire Wardle, “6 Types of Misinformation Circulated This Election Season,” *Columbia Journalism Review*, November 18, 2016, https://www.cjr.org/towcenter/6_types_election_fake_news.php; see also Paul Goble, “Hot Issue – Lies, Damned Lies and Russian Disinformation,” Jamestown Foundation, August 13, 2014, <https://jamestown.org/program/hot-issue-lies-damned-lies-and-russian-disinformation>. As another example, Charleston mass murderer Dylann Roof claimed to have been radicalized after a Google search

- for “black on White crime.” See Rebecca Hersher, “What Happened When Dylann Roof Asked Google for Information About Race?” NPR, January 10, 2017, <https://www.npr.org/sections/thetwo-way/2017/01/10/508363607/what-happened-when-dylann-roof-asked-google-for-information-about-race>.
- 112 For analysis of anti-refugee and anti-migrant disinformation, see Judit Szakács and Éva Bognár, “The Impact of Disinformation Campaigns About Migrants and Minority Groups in the EU,” European Parliament, June 2021, [https://www.europarl.europa.eu/RegData/etudes/IDAN/2021/653641/EXPO_IDA\(2021\)653641_EN.pdf](https://www.europarl.europa.eu/RegData/etudes/IDAN/2021/653641/EXPO_IDA(2021)653641_EN.pdf).
 - 113 To be sure, view count alone does not imply effectiveness. For more about the Our Daily Bread campaign, see “Video Campaign Aims to Unify Poland Through the Power of Bread,” Olga Mecking, NPR, May 21, 2018, <https://www.npr.org/sections/thesalt/2018/05/21/611345277/video-campaign-aims-to-unify-poland-through-the-power-of-bread>.
 - 114 Kevin B. O’Reilly, “Time for Doctors to Take Center Stage in COVID-19 Vaccine Push,” American Medical Association, May 21, 2021, <https://www.ama-assn.org/delivering-care/public-health/time-doctors-take-center-stage-covid-19-vaccine-push>; and Steven Ross Johnson, “Doctors Can Be Key to Higher COVID Vaccination Rates,” U.S. News & World Report, February 28, 2022, <https://www.usnews.com/news/health-news/articles/2022-02-28/primary-care-doctors-can-be-key-to-higher-covid-vaccination-rates>.
 - 115 Filip Viskupić and David L. Wiltse, “The Messenger Matters: Religious Leaders and Overcoming COVID-19 Vaccine Hesitancy,” *Political Science & Politics* 55, no. 3 (2022): <https://www.cambridge.org/core/journals/ps-political-science-and-politics/article/abs/messenger-matters-religious-leaders-and-overcoming-covid19-vaccine-hesitancy/ED93D8BB6C73C8B384986D28B877E284>; and Daniel Estrin and Frank Langfitt, “Religious Leaders Had to Fight Disinformation to Get Their Communities Vaccinated,” NPR, April 23, 2021, <https://www.npr.org/2021/04/23/990281552/religious-leaders-had-to-fight-disinformation-to-get-their-communities-vaccinate>.
 - 116 “The Redirect Method,” Moonshot, accessed March 6, 2023, <https://moonshotteam.com/the-redirect-method/>; and “ADL & Moonshot Partnered to Reduce Extremist Violence During US Presidential Election, Redirected Thousands Towards Safer Content,” Anti-Defamation League, February 1, 2021, <https://www.adl.org/resources/press-release/adl-moonshot-partnered-reduce-extremist-violence-during-us-presidential>.
 - 117 Jacob Davey, Jonathan Birdwell, and Rebecca Skellett, “Counter-conversations: A Model for Direct Engagement With Individuals Showing Signs of Radicalization Online,” Institute for Strategic Dialogue, February 2018, <https://www.isdglobal.org/isd-publications/counter-conversations-a-model-for-direct-engagement-with-individuals-showing-signs-of-radicalisation-online>; and Jacob Davey, Henry Tuck, and Amarnath Amarasingam, “An Imprecise Science: Assessing Interventions for the Prevention, Disengagement and De-radicalisation of Left and Right-Wing Extremists,” Institute for Strategic Dialogue, 2019, <https://www.isdglobal.org/isd-publications/an-imprecise-science-assessing-interventions-for-the-prevention-disengagement-and-de-radicalisation-of-left-and-right-wing-extremists>.
 - 118 On the link between disinformation, hate speech, and hate crimes, consider Jonathan Corpus Ong, “Online Disinformation Against AAPI Communities During the COVID-19 Pandemic,” Carnegie Endowment for International Peace, October 19, 2021, <https://carnegieendowment.org/2021/10/19/online-disinformation-against-aapi-communities-during-covid-19-pandemic-pub-85515>.

- 119 Consider Viskupič and Wiltse, “Messenger Matters”; Scott Bokemper et al., “Testing Persuasive Messaging to Encourage COVID-19 Risk Reduction,” *PLOS ONE* 17 (2022): <https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0264782>; Rupali J. Limaye et al., “Message Testing in India for COVID-19 Vaccine Uptake: What Appeal and What Messenger Are Most Persuasive?,” *Human Vaccines & Immunotherapeutics* 18, no. 6 (2022): <https://www.tandfonline.com/doi/full/10.1080/21645515.2022.2091864>; and Lan Li, Caroline E. Wood, and Patty Kostkova, “Vaccine Hesitancy and Behavior Change Theory-Based Social Media Interventions: A Systematic Review,” *Translational Behavioral Medicine* 12, no. 2 (February 2022): <https://academic.oup.com/tbm/article/12/2/243/6445967>.
- 120 Davey, Tuck, and Amarasingam, “Imprecise Science.”
- 121 Todd C. Helmus and Kurt Klein, “Assessing Outcomes of Online Campaigns Countering Violent Extremism: A Case Study of the Redirect Method,” RAND Corporation, 2018, https://www.rand.org/pubs/research_reports/RR2813.html. The metrics used to evaluate counter-messaging efforts should align with the messenger’s desired outcome, which is not always a direct change in the original speaker’s belief or behavior. Other goals of counter-messaging include influencing passive bystanders to speak out or showing solidarity with a victimized community. See Catherine Buerger, “Why They Do It: Counterspeech Theories of Change,” Dangerous Speech Project, September 26, 2022, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4245211; and Bianca Cepollaro, Maxime Lepoutre, and Robert Mark Simpson, “Counterspeech,” *Philosophy Compass* 18, no. 1 (January 2023): <https://compass.onlinelibrary.wiley.com/doi/full/10.1111/phc3.12890>.
- 122 Limaye et al., “Message Testing in India”; and Li, Wood, and Kostkova, “Vaccine Hesitancy.”
- 123 “Key Takeaways from Civil Society in the Visegrád Region: Fall 2019 Practitioner Convening,” Over Zero, 2019, <https://www.projectoverzero.org/media-and-publications/key-takeaways-from-civil-society-in-the-visegrád-region-fall-2019-practitioner-convening>.
- 124 Consider Viskupič and Wiltse, “Messenger Matters.”
- 125 Davey, Birdwell, and Skellert, “Counter-Conversations”; and Davey, Tuck, and Amarasingam, “Imprecise Science.”
- 126 Dominik Hangartner et al, “Empathy-based Counterspeech Can Reduce Racist Hate Speech in a Social Media Field Experiment,” *PNAS* 118 (2021), <https://www.pnas.org/doi/full/10.1073/pnas.2116310118>.
- 127 Buerger, “Why They Do It.”
- 128 Alysha Ulrich, “Communicating Climate Science in an Era of Misinformation,” *Intersect: The Stanford Journal of Science, Technology, and Society* 16 (2023), <https://ojs.stanford.edu/ojs/index.php/intersect/article/view/2395>.
- 129 Sanchin Banker and Joowon Park, “Evaluating Prosocial COVID-19 Messaging Frames: Evidence from a Field Study on Facebook,” *Judgment and Decision Making* 15 (2023), <https://www.cambridge.org/core/journals/judgment-and-decision-making/article/evaluating-prosocial-covid19-messaging-frames-evidence-from-a-field-study-on-facebook/9EADFB1C6F591AE1A8376C1622FB59D5>.
- 130 Consider Viskupič and Wiltse, “Messenger Matters”; Angela K. Shen et al., “Trusted Messengers and Trusted Messages: The Role for Community-based Organizations in Promoting COVID-19 and Routine Immunizations,” *Vaccine* 41 (2023), <https://www.sciencedirect.com/science/article/pii/S0264410X23001809>; Angela K. Shen et al., “Persuading the ‘Movable Middle’: Characteristics of Effective Messages to Promote Routine and COVID-19 Vaccinations for Adults and Children – The impact of COVID-19 on Beliefs

- and Attitudes,” *Vaccine* 41 (2023), <https://www.sciencedirect.com/science/article/pii/S0264410X2300141X>; and Limaye et al., “Message Testing in India.”
- 131 Benjamin J. Lee, “Informal Countermessaging: The Potential and Perils of Informal Online Countermessaging,” *Studies in Conflict & Terrorism* 42 (2019): <https://www.tandfonline.com/doi/full/10.1080/1057610X.2018.1513697>.
 - 132 Suzanne Smalley, “Collective of Anti-disinformation ‘Elves’ Offer a Bulwark Against Russian Propaganda,” *CyberScoop*, August 9, 2022, <https://cyberscoop.com/collective-anti-disinformation-elves-russian-propaganda>; and Adam Taylor, “With NAFO, Ukraine Turns the Trolls on Russia,” *Washington Post*, September 1, 2022, <https://www.washingtonpost.com/world/2022/09/01/nafo-ukraine-russia>.
 - 133 Davey, Tuck, and Amarasingam, “Imprecise Science.”
 - 134 “Digital Counterterrorism: Fighting Jihadists Online,” Task Force on Terrorism and Ideology, Bipartisan Policy Center, March 2018, <https://bipartisanpolicy.org/download/?file=/wp-content/uploads/2019/03/BPC-National-Security-Digital-Counterterrorism.pdf>; Rita Katz, “The State Department’s Twitter War With ISIS Is Embarrassing,” *Time*, September 16, 2014, <https://time.com/3387065/isis-twitter-war-state-department>; Greg Miller and Scott Higham, “In a Propaganda War Against ISIS, the U.S. Tried to Play by the Enemy’s Rules,” *Washington Post*, May 8, 2015, https://www.washingtonpost.com/world/national-security/in-a-propaganda-war-us-tried-to-play-by-the-enemys-rules/2015/05/08/6eb6b732-e52f-11e4-81ea-0649268f729e_story.html.
 - 135 Samuel Woolley and Katie Joseff, “Demand for Deceit: How the Way We Think Drives Disinformation,” National Endowment for Democracy, January 2020, <https://www.ned.org/wp-content/uploads/2020/01/Demand-for-Deceit.pdf>.
 - 136 Consider Joan Donovan and danah boyd, “Stop the Presses? Moving From Strategic Silence to Strategic Amplification in a Networked Media Ecosystem,” *American Behavioral Scientist* 65, no. 2 (2019): <https://journals.sagepub.com/doi/abs/10.1177/0002764219878229>.
 - 137 “State, Socom Partner to Counter Cyberterrorism,” Simons Center, June 6, 2012, <https://thesimonscenter.org/ia-news/state-socom-partner-to-counter-cyberterrorism>.
 - 138 “Inspection of the Global Engagement Center,” Office of the Inspector General, U.S. Department of State, September 15, 2022, <https://www.oversight.gov/report/DOS/Inspection-Global-Engagement-Center>.
 - 139 “Australia’s COVID-19 Vaccine Information Campaign Begins,” Australian Department of Health and Aged Care, January 27, 2021, <https://www.health.gov.au/ministers/the-hon-greg-hunt-mp/media/australias-covid-19-vaccine-information-campaign-begins>.
 - 140 “Marketing in a Post-Covid Era: Highlights and Insights Report,” CMO Survey, September 2022, https://cmosurvey.org/wp-content/uploads/2022/09/The_CMO_Survey-Highlights_and_Insights_Report-September_2022.pdf.
 - 141 Lee, “Informal Countermessaging.”
 - 142 Hack-and-leak operations might be considered malinformation because the offending material is intended to harm yet is often authentic and factual. Alternatively, such operations could be seen as disinformation to the extent that the term encompasses true but highly misleading information (for example, when crucial context is omitted). For more on this taxonomy, and relevant cybersecurity recommendations, see Wardle and Derakhshan, “Information Disorder.” For further examples of cybersecurity recommendations from this period, see also Jean-Baptiste Jeangène Vilmer, “Successfully Countering Russian Electoral Influence: 15 Lessons Learned

- From the Macron Leaks,” Center for Strategic International Studies, June 2018, https://csis-website-prod.s3.amazonaws.com/s3fs-public/publication/180621_Vilmer_Countering_russian_electoral_influence.pdf; Brattberg and Maurer, “Russian Election Interference”; and Fly, Rosenberger, and Salvo, “The ASD Policy Blueprint.”
- 143 Robby Mook, Matt Rhoades, and Eric Rosenbach, “Cybersecurity Campaign Playbook,” November 2017, <https://www.belfercenter.org/publication/cybersecurity-campaign-playbook>.
 - 144 “Cloudflare for Campaigns: United States,” Cloudflare, accessed April 21, 2023, <https://www.cloudflare.com/campaigns/usa>.
 - 145 “Russian Active Measures Campaigns and Interference in the 2016 U.S. Election, Volume 1: Russian Efforts Against Election Infrastructure With Additional Views,” U.S. Senate Select Committee on Intelligence (16th Congress, Report 116-XX), https://www.intelligence.senate.gov/sites/default/files/documents/Report_Volume1.pdf.
 - 146 Danielle Root, Liz Kennedy, and Michael Sozan, “Election Security in All 50 States,” Center for American Progress, February 12, 2018, <https://www.americanprogress.org/article/election-security-50-states>; Brattberg and Maurer, “Russian Election Interference”; “Statement by Secretary Jeh Johnson on the Designation of Election Infrastructure as a Critical Infrastructure Subsector,” U.S. Department of Homeland Security, January 6, 2017, <https://www.dhs.gov/news/2017/01/06/statement-secretary-johnson-designation-election-infrastructure-critical>; and “Recommendations to Defend America’s Election Infrastructure,” Brennan Center for Justice, October 23, 2019, <https://www.brennancenter.org/our-work/research-reports/recommendations-defend-americas-election-infrastructure>.
 - 147 “Recommendations,” Brennan Center for Justice.
 - 148 “Starting Point: U.S. Election Systems as Critical Infrastructure,” U.S. Election Assistance Commission, accessed March 28, 2023, https://www.eac.gov/sites/default/files/eac_assets/1/6/starting_point_us_election_systems_as_Critical_Infrastructure.pdf; and “DHS Cybersecurity Services Catalog for Election Infrastructure,” U.S. Department of Homeland Security, accessed March 28, 2023, https://www.eac.gov/sites/default/files/eac_assets/1/6/DHS_Cybersecurity_Services_Catalog_for_Election_Infrastructure.pdf.
 - 149 Root, Kennedy, and Sozan, “Election Security”; “Recommendations,” Brennan Center for Justice; and “Elections Infrastructure Information Sharing & Analysis Center,” Center for Internet Security, accessed April 21, 2023, <https://www.cisecurity.org/ei-isac>.
 - 150 Root, Kennedy, and Sozan, “Election Security.”
 - 151 “Federal Cybersecurity Progress Report for Fiscal Year 2022,” U.S. General Services Administration, 2022, <https://www.performance.gov/cyber>. See also “Cross-Sector Cybersecurity Performance Goals,” U.S. Cybersecurity and Infrastructure Security Agency, 2022, https://www.cisa.gov/sites/default/files/publications/2022_00092_CISA_CPG_Report_508c.pdf.
 - 152 “Framework for Improving Critical Infrastructure Cybersecurity,” National Institute of Standards and Technology, April 16, 2018, <https://nvlpubs.nist.gov/nistpubs/cswp/nist.cswp.04162018.pdf>; and “Cybersecurity Solutions for a Riskier World,” ThoughtLab, 2022, <https://thoughtlabgroup.com/cyber-solutions-riskier-world>.
 - 153 See Jeangène Vilmer, “Successfully Countering”; and Brattberg and Maurer, “Russian Election Interference.”
 - 154 For more on pre-bunking, see case studies 3 and 7.

- 155 Dean Jackson and João Guilherme Bastos dos Santos, “A Tale of Two Insurrections: Lessons for Disinformation Research From the Jan. 6 and 8 Attacks,” Lawfare, February 27, 2023, <https://www.lawfaremedia.org/article/tale-two-insurrections-lessons-disinformation-research-jan-6-and-8-attacks-0>.
- 156 See William T. Adler, “To Stop Election-Related Misinformation, Give Election Officials the Resources They Need,” Center for Democracy and Technology, November 13, 2020, <https://cdt.org/insights/to-stop-election-related-misinformation-give-election-officials-the-resources-they-need>.
- 157 William T. Adler and Dhanaraj Thakur, “A Lie Can Travel: Election Disinformation in the United States, Brazil, and France,” Center for Democracy and Technology, December 2021, <https://cdt.org/wp-content/uploads/2021/12/2021-12-13-CDT-KAS-A-Lie-Can-Travel-Election-Disinformation-in-United-States-Brazil-France.pdf>. See also Philip Bump, “The Uncomplicated Reason Brazil Can Count Its Ballots So Quickly,” *Washington Post*, October 31, 2022, <https://www.washingtonpost.com/politics/2022/10/31/brazil-elections-vote-count-united-states>.
- 158 Sam van der Staak, “The Weak Link in Election Security: Europe’s Political Parties,” *Politico*, June 8, 2021, <https://www.politico.eu/article/european-election-security-political-parties-cybersecurity>.
- 159 Brattberg and Maurer, “Russian Election Interference.”
- 160 Isabella Harford, “How Effective Is Security Awareness Training? Not Enough,” TechTarget, April 5, 2022, <https://www.techtarget.com/searchsecurity/feature/How-effective-is-security-awareness-training-Not-enough>.
- 161 Lawrence Norden and Edgardo Cortés, “What Does Election Security Cost?” Brennan Center for Justice, August 15, 2019, <https://www.brennancenter.org/our-work/analysis-opinion/what-does-election-security-cost>.
- 162 Elizabeth Howard et al., “Defending Elections: Federal Funding Needs for State Election Security,” Brennan Center for Justice, July 18, 2019, <https://www.brennancenter.org/our-work/research-reports/defending-elections-federal-funding-needs-state-election-security>.
- 163 Zach Montellaro, “Coronavirus Relief Bill Allocates \$400M for Election Response,” *Politico*, March 26, 2020, <https://www.politico.com/newsletters/morning-score/2020/03/26/coronavirus-relief-bill-allocates-400m-for-election-response-786407>; and “Funding Safe and Secure Elections During the COVID-19 Pandemic,” e.Republic Center for Digital Government, 2020, <https://papers.govtech.com/A-Better-Way-to-Find-New-Jobs-and-Careers-136728.html/Funding-Safe-and-Secure-Elections-During-COVID-19-129933.html>. On the general state of election administration finances, see Tom Scheck, Geoff Hing, Sabby Robinson, and Gracie Stockton, “How Private Money From Facebook’s CEO Saved the 2020 Election,” NPR, December 8, 2020, <https://www.npr.org/2020/12/08/943242106/how-private-money-from-facebooks-ceo-saved-the-2020-election>.
- 164 Rachel Orey, “New Election Security Funding Positive but Misses the Mark,” February 28, 2023, Bipartisan Policy Center, <https://bipartisanpolicy.org/blog/new-election-security-funding>.
- 165 Norden and Cortés, “What Does Election Security Cost?”; and Lawrence Norden, Derek Tisler, and Turquoise Baker, “Estimated Costs for Protecting Election Infrastructure Against Insider Threats,” Brennan Center for Justice, March 7, 2022, <https://www.brennancenter.org/our-work/research-reports/estimated-costs-protecting-election-infrastructure-against-insider>.

- See also Derek Tisler and Lawrence Norden, “Estimated Costs for Protecting Election Workers From Threats of Physical Violence,” Brennan Center for Justice, May 3, 2022, <https://www.brennancenter.org/our-work/research-reports/estimated-costs-protecting-election-workers-threats-physical-violence>.
- 166 Erik Brattberg, “The EU’s Looming Test on Election Interference,” Carnegie Endowment for International Peace, April 18, 2019, <https://carnegieendowment.org/2019/04/18/eu-s-looming-test-on-election-interference-pub-78938>; and Sam van der Staak and Peter Wolf, “Cybersecurity in Elections: Models of Interagency Collaboration,” International IDEA, 2019, <https://www.idea.int/sites/default/files/publications/cybersecurity-in-elections-models-of-interagency-collaboration.pdf>.
 - 167 “Treasury Sanctions Russian Cyber Actors for Interference With the 2016 U.S. Elections and Malicious Cyber-Attacks,” U.S. Department of the Treasury, March 15, 2018, <https://home.treasury.gov/news/press-releases/sm0312>.
 - 168 Devlin Barrett, “U.S. Indicts Two Iranian Hackers Over 2020 Election Disinformation Campaign,” *Washington Post*, November 18, 2021, https://www.washingtonpost.com/national-security/iran-hackers-election-2020-indicted/2021/11/18/605ae112-4898-11ec-b05d-3cb9d96eb495_story.html; and Kevin Breuninger, “DOJ Charges 3 Russians with Running ‘Foreign Influence and Disinformation Network’ in U.S.,” CNBC, April 14 2022, <https://www.cnbc.com/2022/04/14/doj-charges-3-russians-with-running-foreign-influence-and-disinformation-network-in-us.html>.
 - 169 Fried and Polyakova, “Democratic Defense Against Disinformation.”
 - 170 “Foreign Secretary Announces Sanctions on Putin’s Propaganda,” Foreign Ministry of the United Kingdom, March 31, 2022, <https://www.gov.uk/government/news/foreign-secretary-announces-sanctions-on-putins-propaganda--2>.
 - 171 “Report on Foreign Interference in All Democratic Processes in the European Union, Including Disinformation,” European Parliament, 2022, https://www.europarl.europa.eu/doceo/document/A-9-2022-0022_EN.html.
 - 172 “How U.S. Cyber Command, NSA Are Defending Midterm Elections: One Team, One Fight,” U.S. Department of Defense, August 25, 2022, <https://www.defense.gov/News/News-Stories/Article/Article/3138374/how-us-cyber-command-nsa-are-defending-midterm-elections-one-team-one-fight>.
 - 173 On Australia, see “Foreign Influence Transparency Scheme,” Australian Government Attorney-General’s Department, accessed March 13, 2023, <https://www.ag.gov.au/integrity/foreign-influence-transparency-scheme>; Henry Belot, “Malcolm Turnbull Announces Biggest Overhaul of Espionage, Intelligence Laws in Decades,” Australian Broadcasting Corporation, December 4, 2017, <https://www.abc.net.au/news/2017-12-05/turnbull-announces-foreign-interference-laws/9227514>; and Matt Schrader, “Friends and Enemies: A Framework for Understanding Chinese Political Interference in Democratic Countries,” German Marshall Fund, April 2020, <https://securingdemocracy.gmfus.org/wp-content/uploads/2020/05/Friends-and-Enemies-A-Framework-for-Understanding-Chinese-Political-Interference-in-Democratic-Countries.pdf>. On the United Kingdom, see “Foreign Influence Registration Scheme to Make Clandestine Political Activity Illegal,” UK Home Office, October 18, 2022, <https://www.gov.uk/government/news/foreign-influence-registration-scheme-to-make-clandestine-political-activity-illegal>.

- 174 Mark Thompson, “Britain Bans Russian State TV Channel RT,” CNN Business, March 18, 2022, <https://www.cnn.com/2022/03/18/media/uk-bans-russia-rt-tv/index.html>; Foo Yun Chee, “EU Bans RT, Sputnik Over Ukraine Disinformation,” Reuters, March 2 2022, <https://www.reuters.com/world/europe/eu-bans-rt-sputnik-banned-over-ukraine-disinformation-2022-03-02>.
- 175 Keir Giles, “Countering Russian Information Operations in the Age of Social Media,” Council on Foreign Relations, November 21, 2017, <https://www.cfr.org/report/countering-russian-information-operations-age-social-media>.
- 176 For sources on Theresa May, see Paul M. Barrett, Tara Wadhwa, and Dorothee Baumann-Pauly, “Combating Russian Disinformation: The Case for Stepping Up the Fight Online,” NYU Stern Center for Business and Human Rights, July 2018, https://issuu.com/nyusterncenterforbusinessandhumanri/docs/nyu_stern_cbhr_combating_russian_di?e=31640827/63115656; and Rowena Mason, “Theresa May Accuses Russia of Interfering in Elections and Fake News,” *Guardian*, November 14, 2017, <https://www.theguardian.com/politics/2017/nov/13/theresa-may-accuses-russia-of-interfering-in-elections-and-fake-news>. For sources on Finland, see Keir Giles, “Russia’s ‘New’ Tools for Confronting the West: Continuity and Innovation in Moscow’s Exercise of Power,” Chatham House, March 2016, <https://www.chathamhouse.org/sites/default/files/publications/2016-03-russia-new-tools-giles.pdf>; and Aleksi Teivainen, “Report: Haglund Was Quick to Pick Up on Russia’s Disinformation Campaigns,” *Helsinki Times*, March 24, 2016, <https://www.helsinkitimes.fi/finland/finland-news/domestic/13884-report-haglund-was-quick-to-pick-up-on-russia-s-information-campaigns.html>.
- 177 Lizzie Dearden, “Emmanuel Macron Email Leaks ‘Linked to Russian-Backed Hackers Who Attacked Democratic National Committee,” *Independent*, May 6, 2017, <https://www.independent.co.uk/news/world/europe/emmanuel-macron-leaks-hack-en-marche-cyber-attack-russia-dnc-marine-le-pen-election-france-latest-a7721796.html>; Thomas Kent, “Disinformation Response — the Critical Early Hours,” Center for European Policy Analysis, March 31, 2021, <https://cepa.org/article/disinformation-response-the-critical-early-hours>; and Adam Taylor, “U.S. Says Putin Could Use ‘False Flag’ as Excuse for War. Similar Accusations Have Defined Putin’s Career,” February 18, 2022, <https://www.washingtonpost.com/world/2022/02/18/ukraine-putin-false-flag>.
- 178 Andy Greenberg, “US Hackers’ Strike on Russian Trolls Sends a Message—but What Kind?,” *Wired*, February 27, 2019, <https://www.wired.com/story/cyber-command-ira-strike-sends-signal>.
- 179 Jon Bateman, “The Purposes of U.S. Government Public Cyber Attribution,” Carnegie Endowment for International Peace, March 28, 2022, <https://carnegieendowment.org/2022/03/28/purposes-of-u.s.-government-public-cyber-attribution-pub-86696>.
- 180 See Barrett, Wadhwa, and Baumann-Pauly, “Combating Russian Disinformation”; Ben Nimmo and David Agranovich, “Recapping Our 2022 Coordinated Inauthentic Behavior Enforcements,” Meta, December 15, 2022, <https://about.fb.com/news/2022/12/metasp-2022-coordinated-inauthentic-behavior-enforcements>; and “Disinfodex,” Disinfodex, accessed January 27, 2023, <https://disinfodex.org>.
- 181 Gabriel Band, “Sanctions as a Surgical Tool Against Online Foreign Influence,” Lawfare, September 15, 2022, <https://www.lawfaremedia.org/article/sanctions-surgical-tool-against-online-foreign-influence>.
- 182 Bateman, “The Purposes of U.S. Government.”

- 183 Ellen Nakashima, "U.S. Cyber Command Operation Disrupted Internet Access of Russian Troll Factory on Day of 2018 Midterms," *Washington Post*, February 27, 2019, https://www.washingtonpost.com/world/national-security/us-cyber-command-operation-disrupted-internet-access-of-russian-troll-factory-on-day-of-2018-midterms/2019/02/26/1827fc9e-36d6-11e9-af5b-b51b7ff322e9_story.html.
- 184 Sean Lyngaas, "US 'Actively Defending Against Foreign Interference and Influence' in Midterms, Cyber Command Says," CNN, August 25, 2022, <https://www.cnn.com/2022/08/25/politics/election-security-midterms-cyber-command/index.html>.
- 185 Yevgeniy Golovchenko, "Fighting Propaganda With Censorship: A Study of the Ukrainian Ban on Russian Social Media," *Journal of Politics* 84 (2022): <https://www.journals.uchicago.edu/doi/10.1086/716949>.
- 186 Yuyu Chen and David Y. Yang, "The Impact of Media Censorship: 1984 or Brave New World?," *American Economic Review* 109, no. 6 (June 2019): <https://www.aeaweb.org/articles?id=10.1257/aer.20171765>.
- 187 See Elizabeth Dvoskin, Jeremy B. Merrill, and Gerrit De Vynck, "Social Platforms' Bans Muffle Russian State Media Propaganda," *Washington Post*, March 16, 2022, <https://www.washingtonpost.com/technology/2022/03/16/facebook-youtube-russian-bans/>; see also Will Oremus and Cat Zakrzewski, "Big Tech Tried to Quash Russian Propaganda. Russia Found Loopholes," *Washington Post*, August 10, 2022, <https://www.washingtonpost.com/technology/2022/08/10/facebook-twitter-russian-embassy-accounts-propaganda/>.
- 188 Zakary L. Tormala and Richard E. Petty, "Source Credibility and Attitude Certainty: A Metacognitive Analysis of Resistance to Persuasion," *Journal of Consumer Psychology* 14 (2004): <https://www.sciencedirect.com/science/article/abs/pii/S1057740804701692>; Christopher Paul and Miriam Matthews, "The Russian 'Firehose of Falsehood' Propaganda Model: Why It Might Work and Options to Counter It," RAND Corporation, 2016, <https://www.rand.org/pubs/perspectives/PE198.html>.
- 189 Jon Roozenbeek et al., "Psychological Inoculation Improves Resilience Against Misinformation on Social Media," *Science Advances* 8 (2022); and Laura Garcia and Tommy Shane, "A Guide to Prebunking: A Promising Way to Inoculate Against Misinformation," First Draft, June 29, 2021, <https://firstdraftnews.org/articles/a-guide-to-prebunking-a-promising-way-to-inoculate-against-misinformation>.
- 190 Nyhan, "Why the Backfire Effect."
- 191 Greg Myre, "A 'Perception Hack': When Public Reaction Exceeds the Actual Hack," NPR, November 1, 2020, <https://www.npr.org/2020/11/01/929101685/a-perception-hack-when-public-reaction-exceeds-the-actual-hack>.
- 192 Carissa Goodwin and Dean Jackson, "Global Perspectives on Influence Operations Investigations: Shared Challenges, Unequal Resources," Carnegie Endowment for International Peace, February 9, 2022, <https://carnegieendowment.org/2022/02/09/global-perspectives-on-influence-operations-investigations-shared-challenges-unequal-resources-pub-86396>.
- 193 Consider Nimmo and Agranovich, "Recapping Our 2022"; and Goodwin and Jackson, "Global Perspectives."
- 194 Karoun Demirjian and Devlin Barrett, "Obama Team's Response to Russian Election Interference Fell Short, Senate Report Says," *Washington Post*, February 6, 2020, https://www.washingtonpost.com/national-security/obama-teams-response-to-russian-election-interference-fell-short-senate-report-says/2020/02/06/93c2fdac-48f2-11ea-9164-d3154ad8a5cd_story.html.

- 195 Philip Ewing, “FACT CHECK: Why Didn’t Obama Stop Russia’s Election Interference In 2016?” NPR, February 21, 2018, <https://www.npr.org/2018/02/21/587614043/fact-check-why-didnt-obama-stop-russia-s-election-interference-in-2016>.
- 196 “Russia Blocks BBC Website, Says It’s Only Beginning of Its Response,” Reuters, March 16, 2022, <https://www.reuters.com/world/russia-blocks-bbc-website-says-its-only-beginning-its-response-2022-03-16/>
- 197 Joseph Bodnar, “RT en Español Won’t Stay Off YouTube,” German Marshall Fund, March 8, 2023, <https://securingdemocracy.gmfus.org/rt-en-espanol-wont-stay-off-youtube>.
- 198 The scope of this paper was informed by the European Union’s 2022 Strengthened Code of Practice on Disinformation, which covers issues such as fake account creation, bot-driven amplification, and other forms of platform manipulation under “integrity of services.” See “2022 Strengthened Code of Practice on Disinformation,” European Commission, June 16, 2022, <https://digital-strategy.ec.europa.eu/en/library/2022-strengthened-code-practice-disinformation>.
- 199 See Jan Rydzak, “The Stalled Machines of Transparency Reporting,” November 29, 2023, Carnegie Endowment for International Peace, <https://carnegieendowment.org/2023/11/29/stalled-machines-of-transparency-reporting-pub-91085>.
- 200 See “Inauthentic Behavior,” Meta, accessed January 27, 2023, <https://transparency.fb.com/policies/community-standards/inauthentic-behavior>; “Platform Manipulation and Spam Policy,” X, March 2023, <https://help.twitter.com/en/rules-and-policies/platform-manipulation>; and “Civic Integrity Policy,” X, August 2023, <https://help.twitter.com/en/rules-and-policies/election-integrity-policy>.
- 201 For more on how different platform policies address influence operations, see Jon Bateman, Natalie Thompson, and Victoria Smith, “How Social Media Platforms’ Community Standards Address Influence Operations,” April 1, 2021, <https://carnegieendowment.org/2021/04/01/how-social-media-platforms-community-standards-address-influence-operations-pub-84201>.
- 202 “Inauthentic Behavior,” Meta; evelyn douek, “What Does ‘Coordinated Inauthentic Behavior’ Actually Mean?,” Slate, July 2, 2020, <https://slate.com/technology/2020/07/coordinated-inauthentic-behavior-facebook-twitter.html>; and “Election Integrity,” TikTok, accessed January 27, 2023, <https://www.tiktok.com/safety/en-sg/election-integrity>.
- 203 Vijaya Gadde and Yoel Roth, “Enabling Further Research of Information Operations on Twitter,” Twitter, October 17, 2021, https://blog.twitter.com/en_us/topics/company/2018/enabling-further-research-of-information-operations-on-twitter; and “Platform Manipulation and Spam Policy,” X.
- 204 Shane Huntley, “TAG Bulletin: Q3 2022,” Google, October 26, 2022, <https://blog.google/threat-analysis-group/tag-bulletin-q3-2022>.
- 205 Carnegie analysis of “Disinfodex,” Disinfodex.
- 206 “2022 Strengthened Code,” European Commission.
- 207 Samantha Bradshaw, Hannah Bailey, and Philip N. Howard, “Industrialized Disinformation: 2020 Global Inventory of Organized Social Media Manipulation,” Oxford Internet Institute, 2021, <https://demotech.oii.ox.ac.uk/research/posts/industrialized-disinformation>.
- 208 “Threat Report: The State of Influence Operations 2017-2020,” Meta, May 2021, <https://about.fb.com/wp-content/uploads/2021/05/IO-Threat-Report-May-20-2021.pdf>.
- 209 See Camille François and evelyn douek, “The Accidental Origins, Underappreciated Limits, and Enduring Promises of Platform Transparency Reporting about Information Operations,”

- Journal of Online Trust & Safety* 1 (2021): <https://tsjournal.org/index.php/jots/article/view/17>; and Samantha Lai, Naomi Shiffman, Alicia Wanless, "Operational Reporting by Online Services: A Proposed Framework," Carnegie Endowment for International Peace, May 18, 2023, <https://carnegieendowment.org/2023/05/18/operational-reporting-by-online-services-proposed-framework-pub-89776>.
- 210 Joseph Menn, Elizabeth Dwoskin, and Cat Zakrzewski, "Former Security Chief Claims Twitter Buried 'Egregious Deficiencies,'" *Washington Post*, August 23, 2022, <https://www.washingtonpost.com/technology/interactive/2022/twitter-whistleblower-sec-spam>.
- 211 "Foreign Threats to the 2020 US Federal Elections," U.S. National Intelligence Council, March 10, 2021, <https://www.dni.gov/files/ODNI/documents/assessments/ICA-declass-16MAR21.pdf>.
- 212 Ben Nimmo, "Assessing the Impact of Influence Operations Through the Breakout Scale," Carnegie Endowment for International Peace, October 25, 2022, <https://carnegieendowment.org/2022/10/25/perspectives-for-influence-operations-investigators-pub-88208#breakout>.
- 213 "Foreign Threats to the 2020 US Federal Elections," U.S. National Intelligence Council.
- 214 Bateman, Hickok, Courchesne, Thange, and Shapiro, "Measuring the Effects."
- 215 Joshua A. Tucker, "The Limited Room for Russian Troll Influence in 2016," *Lawfare*, October 27, 2020, <https://www.lawfaremedia.org/article/limited-room-russian-troll-influence-2016>; see also Gregory Eady et al., "Exposure to the Russian Internet Research Agency Foreign Influence Campaign on Twitter in the 2016 US Election and Its Relationship to Attitudes and Voting Behavior," *Nature Communications* 14 (2023), <https://www.nature.com/articles/s41467-022-35576-9>.
- 216 Soroush Vosoughi, Deb Roy, and Sinan Aral, "The Spread of True and False News Online," *Science* 359 (2018): <https://www.science.org/doi/10.1126/science.aap9559>.
- 217 Jackson and Dos Santos, "Tale of Two Insurrections."
- 218 Nathaniel Gleicher, "Removing New Types of Harmful Networks," *Meta*, September 16, 2021, <https://about.fb.com/news/2021/09/removing-new-types-of-harmful-networks>; and "An Update to How We Address Movements and Organizations Tied to Violence," *Meta*, October 17, 2022, <https://about.fb.com/news/2020/08/addressing-movements-and-organizations-tied-to-violence>.
- 219 See Tamar Mitts, "Banned: How Deplatforming Extremists Mobilizes Hate in the Dark Corners of the Internet," Tel Aviv University, October 18, 2021, https://social-sciences.tau.ac.il/sites/socsci.tau.ac.il/files/media_server/Governemnt/Mitts_Digital_CT_Banned.pdf; and Tamar Mitts, Nilima Pisharody, and Jacob Shapiro, "Removal of Anti-vaccine Content Impacts Social Media Discourse," (paper presented at the 14th ACM Web Science Conference, Barcelona, June 26–29, 2022), <https://dl.acm.org/doi/10.1145/3501247.3531548>.
- 220 Joseph Menn, Elizabeth Dwoskin, and Cat Zakrzewski, "Twitter Can't Afford to Be One of the World's Most Influential Websites," *Washington Post*, September 4, 2022, <https://www.washingtonpost.com/technology/2022/09/04/twitter-mudge-alethea-resources>.
- 221 Menn, Dwoskin, and Zakrzewski, "Twitter Can't Afford"; and Menn, Dwoskin, and Zakrzewski, "Former Security Chief."
- 222 Menn, Dwoskin, and Zakrzewski, "Twitter Can't Afford."

- 223 Ben Popper, “Facebook’s Business Is Booming, but It Says Preventing Abuse Will Cut Into Future Profits,” *The Verge*, November 1, 2017, <https://www.theverge.com/2017/11/1/16593812/facebook-earnings-q3-third-quarter-2017>; and Matthew Mahoney, “Twitter and Its Fight for Profitability,” *Michigan Journal of Economics*, December 21, 2022, <https://sites.lsa.umich.edu/mje/2022/12/21/twitter-and-its-fight-for-profitability>.
- 224 “Current State Assessment,” Althea Group, as attached to Menn, Dwoskin, and Zakrzewski, “Former Security Chief.”
- 225 See Victoria Smith and Jon Bateman, “Best Practice Guidance for Influence Operations Research: Survey Shows Needs and Challenges,” Carnegie Endowment for International Peace, August 2, 2022, <https://carnegieendowment.org/2022/08/02/best-practice-guidance-for-influence-operations-research-survey-shows-needs-and-challenges-pub-87601>; Dean Jackson, “Influence Operations Researchers Want Guidance on Best Practice, But What Does That Mean?” Carnegie Endowment for International Peace, December 5, 2022, <https://carnegieendowment.org/2022/12/05/influence-operations-researchers-want-guidance-on-best-practice-but-what-does-that-mean-pub-88517>; and Victoria Smith and Natalie Thompson, “Survey on Countering Influence Operations Highlights Steep Challenges, Great Opportunities,” Carnegie Endowment for International Peace, December 7, 2020, <https://carnegieendowment.org/2020/12/07/survey-on-countering-influence-operations-highlights-steep-challenges-great-opportunities-pub-83370>.
- 226 See Odanga Madung, “Jack Dorsey’s Twitter Failed African Countries,” *Wired*, December 3, 2021, <https://www.wired.com/story/jack-dorseys-twitter-failed-african-countries>; and Goodwin and Jackson, “Global Perspectives.”
- 227 Consider Nathaniel Persily, “The 2016 U.S. Election: Can Democracy Survive the Internet?” *Journal of Democracy* 28 (April 2017): <https://www.journalofdemocracy.org/articles/the-2016-u-s-election-can-democracy-survive-the-internet>.
- 228 One commentator, for example, called data privacy laws an “elegant arrow in the quiver of responses to online disinformation” that can “protect against foreign and homegrown trolls alike.” See Alex Campbell, “How Data Privacy Laws Can Fight Fake News,” *Just Security*, August 15, 2019, <https://www.justsecurity.org/65795/how-data-privacy-laws-can-fight-fake-news>. See also Balázs Bodó, Natali Helberger, and Claes de Vreese, “Political Micro-targeting: A Manchurian Candidate or Just a Dark Horse?” *Internet Policy Review* 6 (2017): <https://policyreview.info/articles/analysis/political-micro-targeting-manchurian-candidate-or-just-dark-horse>; and Dipayan Ghosh and Ben Scott, “Digital Deceit: The Technologies Behind Precision Propaganda on the Internet,” *New America*, January 23, 2018, <https://www.newamerica.org/pit/policy-papers/digitaldeceit>.
- 229 “Karl Rove: The Architect,” *PBS Frontline*, April 12, 2005, <https://www.pbs.org/wgbh/pages/frontline/shows/architect>. For more discussion on the definition of micro-targeting, see Tom Dobber, Damian Trilling, Natali Helberger, and Claes de Vreese, “Effects of an Issue-Based Microtargeting Campaign: A Small-Scale Field Experiment in a Multi-party Setting,” *Information Society* 39 (2022): <https://www.tandfonline.com/doi/full/10.1080/01972243.2022.2134240>.
- 230 “Personal Data: Political Persuasion,” *Tactical Tech*, March 2019, https://cdn.ttc.io/s/tacticaltech.org/methods_guidebook_A4_spread_web_Ed2.pdf.
- 231 Nathalie Maréchal and Ellery Roberts Biddle, “It’s Not Just the Content, It’s the Business Model: Democracy’s Online Speech Challenge,” *New America*, March 17, 2020, <https://www.newamerica.org/oti/reports/its-not-just-content-its-business-model>; see also Anthony

- Nadler, Matthew Crain, and Joan Donovan, “Weaponizing the Digital Influence Machine: The Political Perils of Online Ad Tech,” *Data & Society*, October 2018, <https://datasociety.net/library/weaponizing-the-digital-influence-machine>.
- 232 For one example, see Karen Kornbluh, “Could Europe’s New Data Protection Regulation Curb Online Disinformation?,” *Council on Foreign Relations*, February 20, 2018, <https://www.cfr.org/blog/could-europes-new-data-protection-regulation-curb-online-disinformation>.
- 233 “Who Does the Data Protection Law Apply To?” European Commission, accessed December 10, 2023, https://commission.europa.eu/law/law-topic/data-protection/reform/rules-business-and-organisations/application-regulation/who-does-data-protection-law-apply_en.
- 234 Matt Burgess, “What Is GDPR? The Summary Guide to GDPR Compliance in the UK,” *Wired UK*, March 24, 2020, <https://www.wired.co.uk/article/what-is-gdpr-uk-eu-legislation-compliance-summary-fines-2018>; “Anonymisation and Pseudonymisation,” University College London, accessed April 5, 2023, <https://www.ucl.ac.uk/data-protection/guidance-staff-students-and-researchers/practical-data-protection-guidance-notices/anonymisation-and>; and Joshua Gresham, “Is Encrypted Data Personal Data Under the GDPR?,” *International Association of Privacy Professionals*, March 6, 2019, <https://iapp.org/news/a/is-encrypted-data-personal-data-under-the-gdpr>.
- 235 Zachey Klinger, “A Federal Data Privacy Law May Be the Best Tool to Combat Online Disinformation,” *Tech Policy Press*, April 16, 2021, <https://techpolicy.press/a-federal-data-privacy-law-may-be-the-best-tool-to-combat-online-disinformation>; see also Campbell, “How Data Privacy Laws.”
- 236 “GDPR Loopholes Facilitate Data Exploitation by Political Parties,” *Privacy International*, April 30, 2019, <https://privacyinternational.org/news-analysis/2836/gdpr-loopholes-facilitate-data-exploitation-political-parties>.
- 237 “The Digital Services Act Package,” European Commission, accessed April 5, 2023, <https://digital-strategy.ec.europa.eu/en/policies/digital-services-act-package>; and “Transparency and Targeting of Political Advertising: EU Co-legislators Strike Deal on New Regulation,” European Council, November 7, 2023, <https://www.consilium.europa.eu/en/press/press-releases/2023/11/07/transparency-and-targeting-of-political-advertising-eu-co-legislators-strike-deal-on-new-regulation>.
- 238 Clothilde Goujard, “European Parliament Votes to Transform Online Political Campaigning,” *Politico Europe*, February 2, 2023, <https://www.politico.eu/article/european-parliament-approves-its-position-on-political-advertising-law>; and Mark Scott and Clothilde Goujard, “5 Things You Need to Know About Europe’s Political Ad Rules,” *Politico Europe*, November 23, 2021, <https://www.politico.eu/article/political-ads-europe-facebook-google>.
- 239 “Online Targeting for Political Advertising: Stricter Rules Are Necessary,” Office of the European Data Protection Supervisor, January 20, 2022, https://edps.europa.eu/press-publications/press-news/press-releases/2022/online-targeting-political-advertising-stricter_en; and Goujard, “European Parliament Votes.”
- 240 Kate Conger, “Twitter Will Ban All Political Ads, C.E.O. Jack Dorsey Says,” *New York Times*, October 30, 2019, <https://www.nytimes.com/2019/10/30/technology/twitter-political-ads-ban.html>; Kate Conger, “What Ads Are Political? Twitter Struggles With a Definition,” *New York Times*, November 15, 2019, <https://www.nytimes.com/2019/11/15/technology/twitter-political-ad-policy.html>; and Kate Conger, “Twitter to Relax Ban on Political Ads,” *New York Times*, January 3, 2023, <https://www.nytimes.com/2023/01/03/technology/twitter-political-ads.html>.

- 241 Makena Kelly, “Google Issues Harsh New Restrictions on Political Ad Targeting,” *The Verge*, November 20, 2019, <https://www.theverge.com/2019/11/20/20975054/google-advertising-political-rules-twitter-ban-election-uk-general-2020>; and Kate Kaye, “‘We Will Not Build Alternate Identifiers’: In Drastic Shift, Google Will End Behavioral Targeting, Profile-Building in Its Ad Products,” *Digiday*, March 3, 2021, <https://digiday.com/media/we-will-not-build-alternate-identifiers-in-drastic-shift-google-will-end-behavioral-targeting-profile-building-in-its-ad-products>.
- 242 Jeff Horwitz, “Facebook Parent Meta Limits Ad Targeting for Politics and Other Sensitive Issues,” *Wall Street Journal*, November 9, 2014, <https://www.wsj.com/articles/facebook-parent-meta-bans-targeting-for-political-ads-11636488053>.
- 243 Kliger, “Federal Data Privacy Law”; Iva Nenadić, “Unpacking the “European Approach” to Tackling Challenges of Disinformation and Political Manipulation,” *Internet Policy Review*, December 31, 2019, <https://policyreview.info/articles/analysis/unpacking-european-approach-tackling-challenges-disinformation-and-political>; and Razieh Nokhbeh Zaeem and K. Suzanne Barber, “The Effect of the GDPR on Privacy Policies: Recent Progress and Future Promise,” *ACM Transactions and Management Information Systems* 12 (2020): <https://dl.acm.org/doi/abs/10.1145/3389685>.
- 244 Jeremy Kahn, Stephanie Bodoni, and Stefan Nicola, “It’ll Cost Billions for Companies to Comply With Europe’s New Data Law,” *Bloomberg*, March 22, 2018, <https://www.bloomberg.com/news/articles/2018-03-22/it-ll-cost-billions-for-companies-to-comply-with-europe-s-new-data-law>; and Samuel Goldberg, Garrett Johnson, and Scott Shriver, “Regulating Privacy Online: An Economic Evaluation of the GDPR,” Law & Economics Center at George Mason University Scalia Law School, Research Paper Series no. 22-025, November 17, 2022, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3421731.
- 245 Sara Lebow, “Worldwide Digital Ad Spend Will Top \$600 Billion This Year,” *Insider Intelligence*, January 31, 2023, <https://www.insiderintelligence.com/content/worldwide-digital-ad-spend-will-top-600-billion-this-year>.
- 246 Brahim Zarouali, Tom Dobber, Guy de Pauw, and Claes de Vreese, “Using a Personality-Profiling Algorithm to Investigate Political Microtargeting: Assessing the Persuasion Effects of Personality-Tailored Ads on Social Media,” *Communication Research* 49 (2020): <https://journals.sagepub.com/doi/full/10.1177/0093650220961965>; see also Bodó, Helberger, and de Vreese, “Political Micro-targeting.”
- 247 For example, see Bernstein, “Bad News.”
- 248 Bodó, Helberger, and de Vreese, “Political Micro-targeting.” Consider Peter J. Danaher, “Optimal Microtargeting of Advertising,” *Journal of Marketing Research* 60 (2022): <https://journals.sagepub.com/doi/full/10.1177/0022243722116034>; see also Brian Resnick, “Cambridge Analytica’s ‘Psychographic Microtargeting’: What’s Bullshit and What’s Legit,” *Vox*, March 26, 2018, <https://www.vox.com/science-and-health/2018/3/23/17152564/cambridge-analytica-psychographic-microtargeting-what>.
- 249 Dobber, Trilling, Helberger, and de Vreese, “Effects of Issue-Based Microtargeting”; and Bodó, Helberger, and de Vreese, “Political Micro-targeting.”
- 250 Dobber, Trilling, Helberger, and de Vreese, “Effects of Issue-Based Microtargeting.”
- 251 Dobber, Trilling, Helberger, and de Vreese, “Effects of Issue-Based Microtargeting”; see also Resnick, “Cambridge Analytica’s ‘Psychographic Microtargeting.’”
- 252 For more on Cambridge Analytica, see Resnick, “Cambridge Analytica’s ‘Psychographic Microtargeting.’”

- 253 Zarouali, Dobber, de Pauw, and de Vreese, “Using a Personality-profiling Algorithm”; and Lennart J. Krotzek, “Inside the Voter’s Mind: The Effect of Psychometric Microtargeting on Feelings Toward and Propensity to Vote for a Candidate,” *International Journal of Communication* 13 (2019): <https://ijoc.org/index.php/ijoc/article/view/9605/2742>.
- 254 Goldberg, Johnson, and Shriver, “Regulating Privacy Online.”
- 255 Giorgio Presidente and Carl Benedikt Frey, “The GDPR Effect: How Data Privacy Regulation Shaped Firm Performance Globally,” *VoxEU* (blog), Centre for Economic Policy Research, March 10, 2022, <https://cepr.org/voxeu/columns/gdpr-effect-how-data-privacy-regulation-shaped-firm-performance-globally>; see also Pete Swabey, “GDPR Cost Businesses 8% of Their Profits, According to a New Estimate,” *Tech Monitor*, March 11, 2022, <https://techmonitor.ai/policy/privacy-and-data-protection/gdpr-cost-businesses-8-of-their-profits-according-to-a-new-estimate>.
- 256 Fredric D. Bellamy, “U.S. Data Privacy Laws to Enter New Era in 2023,” Reuters, January 12, 2023, <https://www.reuters.com/legal/legalindustry/us-data-privacy-laws-enter-new-era-2023-2023-01-12>; “Data Privacy Laws by State: Comparison Charts,” Bloomberg Law, July 11, 2023, <https://pro.bloomberglaw.com/brief/privacy-laws-us-vs-eu-gdpr>; Jonathan Keane, “From California to Brazil: Europe’s Privacy Laws Have Created a Recipe for the World,” CNBC, April 8, 2021, <https://www.cnbc.com/2021/04/08/from-california-to-brazil-gdpr-has-created-recipe-for-the-world.html>.
- 257 Goldberg, Johnson, and Shriver, “Regulating Privacy Online.”
- 258 Kahn, Bodoni, and Nicola, “It’ll Cost Billions”; and Goldberg, Johnson, and Shriver, “Regulating Privacy Online.”
- 259 Patience Haggin, Keach Hagey, and Sam Schechner, “Apple’s Privacy Change Will Hit Facebook’s Core Ad Business. Here’s How,” *Wall Street Journal*, January 9, 2021, <https://www.wsj.com/articles/apples-privacy-change-will-hit-facebooks-core-ad-business-heres-how-11611938750>; and Peter Kafka, “Apple Broke Facebook’s Ad Machine. Who’s Going to Fix It?” *Vox*, February 14, 2022, <https://www.vox.com/recode/22929715/facebook-apple-ads-meta-privacy>.
- 260 See Alan McQuinn and Daniel Castro, “The Costs of an Unnecessarily Stringent Federal Data Privacy Law,” Information Technology & Innovation Foundation, August 5, 2019, <https://itif.org/publications/2019/08/05/costs-unnecessarily-stringent-federal-data-privacy-law>.
- 261 Kilger, “Federal Data Privacy Law.”
- 262 Megan Graham, “Advertising Market Keeps Growing Much Faster Than Expected, Forecasters Say,” *Wall Street Journal*, December 6, 2021, <https://www.wsj.com/articles/advertising-market-keeps-growing-much-faster-than-expected-forecasters-say-11638784800>; and “The Online-ad Industry Is Being Shaken Up,” *Economist*, July 28, 2022, <https://www.economist.com/business/2022/07/28/the-online-ad-industry-is-being-shaken-up>.
- 263 Goujard, “European Parliament Votes.”
- 264 Tarleton Gillespie, “Do Not Recommend? Reduction as a Form of Content Moderation,” *Social Media + Society* 8 (2022): <https://journals.sagepub.com/doi/full/10.1177/20563051221117552>.
- 265 Maréchal and Biddle, “It’s Not Just the Content.”
- 266 Vosoughi, Roy, and Aral, “Spread of True and False News.”
- 267 Geoff Nunberg, “‘Disinformation’ Is the Word of the Year—and a Sign of What’s to Come,” NPR, December 30, 2019, <https://www.npr.org/2019/12/30/790144099/disinformation-is-the-word-of-the-year-and-a-sign-of-what-s-to-come>.

- 268 Jeff Horwitz and Deepa Seetharaman, “Facebook Executives Shut Down Efforts to Make the Site Less Divisive,” *Wall Street Journal*, May 26, 2020, <https://www.wsj.com/articles/facebook-knows-it-encourages-division-top-executives-nixed-solutions-11590507499>; and Loveday Morris, “In Poland’s Politics, a ‘Social Civil War’ Brewed as Facebook Rewarded Online Anger,” *Washington Post*, October 27, 2021, <https://www.washingtonpost.com/world/2021/10/27/poland-facebook-algorithm>.
- 269 Samanth Subramanian, “The Macedonian Teens Who Mastered Fake News,” *Wired*, February 15, 2017, <https://www.wired.com/2017/02/veles-macedonia-fake-news>.
- 270 Consider Swati Chaturvedi, *I Am a Troll: Inside the Secret World of the BJP’s Digital Army* (Juggernaut Books, 2016); and Jonathan Corpus Ong and Jason Vincent A. Cabañes, “Architects of Networked Disinformation: Behind the Scenes of Troll Accounts and Fake News Production in the Philippines,” University of Massachusetts Amherst, Communications Department Faculty Publication Series, 2018, https://scholarworks.umass.edu/cgi/viewcontent.cgi?article=1075&context=communication_faculty_pubs.
- 271 Renée DiResta, “Free Speech Is Not the Same as Free Reach,” *Wired*, August 30, 2018, <https://www.wired.com/story/free-speech-is-not-the-same-as-free-reach>.
- 272 Jennifer McCoy, Benjamin Press, Murat Somer, and Ozlem Tuncel, “Reducing Pernicious Polarization: A Comparative Historical Analysis of Depolarization,” Carnegie Endowment for International Peace, May 5, 2022, <https://carnegieendowment.org/2022/05/05/reducing-pernicious-polarization-comparative-historical-analysis-of-depolarization-pub-87034>.
- 273 A review by New York University’s Stern Center for Business and Human Rights found that while “Facebook, Twitter, and YouTube are not the original or main cause of rising U.S. political polarization . . . use of those platforms intensifies divisiveness and thus contributes to its corrosive consequences.” While observers sometimes attribute increased polarization to social media echo chambers, scholars note that most users see a diverse range of news sources. In fact, online heterogeneity might be related to *increased* polarization: one study based on Twitter data suggests that “it is not isolation from opposing views that drives polarization but precisely the fact that digital media” collapses the cross-cutting conflicts that occur in local interactions, thereby hardening partisan identities and deepening division. (Other, non-digital sources of national news are believed to promote similar dynamics.) And the evidence is not all in one direction: in a study on Bosnia and Herzegovina, abstinence from Facebook use was associated with *lower* regard for ethnic outgroups, suggesting Facebook could have a depolarizing effect on users. See Paul Barrett, Justin Hendrix, and J. Grant Sims, “Fueling the Fire: How Social Media Intensifies U.S. Political Polarization — And What Can Be Done About It,” NYU Stern Center for Business and Human Rights, September 13, 2021, https://bhr.stern.nyu.edu/polarization-report-page?_ga=2.126094349.1087885125.1705371436-402766718.1705371436; Andrew Guess, Brendan Nyhan, Benjamin Lyons, and Jason Reifler, “Avoiding the Echo Chamber About Echo Chambers,” Knight Foundation, 2018, https://kf-site-production.s3.amazonaws.com/media_elements/files/000/000/133/original/Topos_KF_White-Paper_Nyhan_V1.pdf; Petter Törnberg, “How Digital Media Drive Affective Polarization Through Partisan Sorting,” *PNAS* 119 (2022): <https://www.pnas.org/doi/10.1073/pnas.2207159119>; Darr, Hitt, and Dunaway, “Newspaper Closures Polarize”; and Nejla Ašimović, Jonathan Nagler, Richard Bonneau, and Joshua A. Tucker, “Testing the Effects of Facebook Usage in an Ethnically Polarized Setting,” *PNAS* 118 (2021): <https://www.pnas.org/doi/10.1073/pnas.2022819118>.

- 274 Annie Y. Chen, Brendan Nyhan, Jason Reifler, Ronald E. Robertson, and Christo Wilson, "Subscriptions and External Links Help Drive Resentful Users to Alternative and Extremist YouTube Videos," *Science Advances* 9 (2023): <https://www.science.org/doi/10.1126/sciadv.add8080>.
- 275 As discussed in case study 1, polarization and its effect on spreading disinformation is not symmetrical across the political spectrum. In most (but not all) political contexts, the right side of the political spectrum has moved further from the mainstream both ideologically and in its media diet, creating more appetite for partisan disinformation. See Joshua A. Tucker et al., "Social Media, Political Polarization, and Political Disinformation: A Review of the Scientific Literature," Hewlett Foundation, March 2018, https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3144139.
- 276 Meta refers to "prevalence" of content on its platform as a measure of how many users *see* a specific piece of content, a metric which "assumes that the impact caused by violating content is proportional to the number of times that content is viewed." See "Prevalence," Meta, November 18, 2022, <https://transparency.fb.com/policies/improving/prevalence-metric/>. Some companies may contest that "engagement" is the primary motive behind platform algorithms, or they may point to certain forms of engagement as positive for society and therefore an acceptable optimization goal for curation algorithms. Meta, for instance, says that it promotes "meaningful interactions" between users instead of mere clicks. Engagement is used here as a shorthand for the broad goal of motivating users to continue using a service, especially for longer periods of time. See Adam Mosseri, "News Feed FYI: Bringing People Closer Together," Meta, January 11, 2018, <https://www.facebook.com/business/news/news-feed-fyi-bringing-people-closer-together>.
- 277 Shannon Bond and Bobby Allyn, "How the 'Stop the Steal' Movement Outwitted Facebook Ahead of the Jan. 6 Insurrection," NPR, October 22, 2021, <https://www.npr.org/2021/10/22/1048543513/facebook-groups-jan-6-insurrection>.
- 278 Robert Gorwa, Reuben Binns, and Christian Katzenbach, "Algorithmic Content Moderation: Technical and Political Challenges in the Automation of Platform Governance," *Big Data & Society* 7 (2020): <https://journals.sagepub.com/doi/10.1177/2053951719897945>.
- 279 Renée DiResta, "Up Next: A Better Recommendation System," *Wired*, April 11, 2018, <https://www.wired.com/story/creating-ethical-recommendation-engines>.
- 280 Chand Rajendra-Nicolucci and Ethan Zuckerman, "An Illustrated Field Guide to Social Media," Knight First Amendment Institute, May 14, 2021, <https://knightcolumbia.org/blog/an-illustrated-field-guide-to-social-media>.
- 281 Aviv Ovadya and Luke Thorburn, "Bridging Systems: Open Problems for Countering Destructive Divisiveness Across Ranking, Recommenders, and Governance?," Knight First Amendment Institute, October 26, 2023, <https://knightcolumbia.org/content/bridging-systems>.
- 282 Joan Donovan, "Shhhh... Combating the Cacophony of Content with Librarians," National Endowment for Democracy, January 2021, <https://www.ned.org/wp-content/uploads/2021/01/Combating-Cacophony-Content-Librarians-Donovan.pdf>.
- 283 See Francis Fukuyama, Barak Richman, and Ashish Goel, "How to Save Democracy from Technology: Ending Big Tech's Information Monopoly," *Foreign Affairs*, November 24, 2020, <https://www.foreignaffairs.com/articles/united-states/2020-11-24/fukuyama-how-save-democracy-technology>; and Francis Fukuyama, Barak Richman, Ashish Goel, Roberta R. Katz, A. Douglas Melamed, and Marietje Schaake, "Middleware for Dominant Digital

Platforms: A Technological Solution to a Threat to Democracy,” Stanford Cyber Policy Center, accessed March 13, 2023, https://fsi-live.s3.us-west-1.amazonaws.com/s3fs-public/cpc-middleware_ff_v2.pdf.

- 284 See “The Future of Platform Power,” various authors, *Journal of Democracy* 32 (2021): <https://www.journalofdemocracy.org/issue/july-2021/>.
- 285 Katharine Miller, “Radical Proposal: Middleware Could Give Consumers Choices Over What They See Online,” Stanford University Human-Centered Artificial Intelligence, October 20, 2021, <https://hai.stanford.edu/news/radical-proposal-middleware-could-give-consumers-choices-over-what-they-see-online>.
- 286 Consider, for example, Ronald E. Robertson, “Uncommon Yet Consequential Online Harms,” *Journal of Online Trust & Safety* 1 (2022): <https://tsjournal.org/index.php/jots/article/view/87>.
- 287 Ronald E. Robertson et al., “Users Choose to Engage With More Partisan News than They Are Exposed to on Google Search,” *Nature* 618 (2022): <https://www.nature.com/articles/s41586-023-06078-5>.
- 288 Adam Satariano, “Meta’s Ad Practices Ruled Illegal Under E.U. Law,” *New York Times*, January 4, 2023, <https://www.nytimes.com/2023/01/04/technology/meta-facebook-eu-gdpr.html>.
- 289 Rajendra-Nicolucci and Zuckerman, “Illustrated Field Guide.”
- 290 Natasha Lomas, “All Hail the New EU Law That Lets Social Media Users Quiet Quit the Algorithm,” TechCrunch, August 25, 2023, <https://techcrunch.com/2023/08/25/quiet-quitting-ai>.
- 291 See, for example, Jon Bateman, “Deepfakes and Synthetic Media in the Financial System: Assessing Threat Scenarios,” Carnegie Endowment for International Peace, July 8, 2020, <https://carnegieendowment.org/2020/07/08/deepfakes-and-synthetic-media-in-financial-system-assessing-threat-scenarios-pub-82237>.
- 292 Jane Wakefield, “Deepfake Presidents Used in Russia-Ukraine War,” BBC, March 18, 2022, <https://www.bbc.com/news/technology-60780142>.
- 293 “C2PA Explainer,” Coalition for Content Provenance and Authenticity, accessed March 13, 2023, <https://c2pa.org/specifications/specifications/1.0/explainer/Explainer.html>.
- 294 Karen Hao, “How Facebook Uses Machine Learning to Detect Fake Accounts,” MIT Technology Review, March 4, 2020, <https://www.technologyreview.com/2020/03/04/905551/how-facebook-uses-machine-learning-to-detect-fake-accounts>.
- 295 Alicia Wanless, “There Is No Getting Ahead of Disinformation Without Moving Past It,” Lawfare, May 8, 2023, <https://www.lawfaremedia.org/article/there-is-no-getting-ahead-of-disinformation-without-moving-past-it>; and Alicia Wanless, “Information Ecology: Using Physical Ecology to Understand Information Struggle,” King’s College London, forthcoming, abstract accessed on December 11, 2023, <https://www.kcl.ac.uk/people/alicia-wanless>.

CARNEGIE ENDOWMENT FOR INTERNATIONAL PEACE

The Carnegie Endowment for International Peace is a unique global network of policy research centers around the world. Our mission, dating back more than a century, is to advance peace through analysis and development of fresh policy ideas and direct engagement and collaboration with decisionmakers in government, business, and civil society. Working together, our centers bring the inestimable benefit of multiple national viewpoints to bilateral, regional, and global issues.

TECHNOLOGY AND INTERNATIONAL AFFAIRS PROGRAM

The Technology and International Affairs Program develops insights to address the governance challenges and large-scale risks of new technologies. Our experts identify actionable best practices and incentives for industry and government leaders on artificial intelligence, cyber threats, cloud security, countering influence operations, reducing the risk of biotechnologies, and ensuring global digital inclusion.



CarnegieEndowment.org