Berkeley
UNIVERSITY OF CALIFORNIA
BERKELEY ROUNDTABLE ON
THE INTERNATIONAL ECONOMY

BRIE Working Paper
2021-7

Technology Solutions for Disinformation

Mark Nitzberg with Camille Carlton

Technology Solutions to Promote Information Integrity

Mark Nitzberg
Director of Technology Research,
Berkeley Roundtable on the International Economy (BRIE);
Executive Director,
Center for Human-Compatible Artificial Intelligence (CHAI);
Head of Strategic Outreach,
Berkeley AI Research (BAIR)
University of California, Berkeley
nitz@berkeley.edu

with

Camille Carlton
Policy, Communications, and Research, Tech Governance,
Berkeley Roundtable on the International Economy (BRIE);
University of California, Berkeley
camillecarlton@berkeley.edu

Abstract In response to the disinformation crisis, new policies, practices, and regulations will need corresponding practical technology solutions. This paper outlines five key functions that technology can serve to restore information integrity: measure and track offline harms from online information, enable responsible algorithm design and review, label content, throttle the spread of harmful content, and track-and-trace online mis- and disinformation to its sources. Fortunately, many of the tools are available and in use. However, they need to be integrated into a holistic system that begins with recognizing mis- and disinformation and ends with stopping the problem at its source.

Introduction

Algorithmic spread has weakened information integrity[1] to the point of national crisis. Big Tech firms[2] determine what half the people on earth read, see, and hear online every day, using content selection algorithms tuned to meet a commercial mandate, without a countervailing social mandate. The combination of scale and automation, often with the support of malicious actors, threatens the foundations of our nation's civil society and democratic government.

This paper outlines key technology solutions needed for the practical implementation of a plan to restore information integrity. They include the technical means to measure and track offline harms from online information, to enable responsible algorithm design and review, to label content, to throttle the spread of harmful content, and to track-and-trace online mis- and disinformation to its sources. As private, public, and civil hand-wringing turns to policy and regulation, we believe these are the technology elements needed to restore information integrity. In a world where technology is interposed in every human activity, policy goals are irrelevant without the computer code to put them into practice.

Background

Let us review the consequences and implications of today's unregulated technology ecosystem before turning to solutions. Pervasive Internet infrastructure, mobile phones, and digital platforms have combined to shift the delivery and consumption of news to digital mediums. The increase in scale and reach of information has placed authoritative stalwarts of news such as the New York Times and the Wall Street Journal on a par with a vast array of sources that vary widely in quality and verifiability. News distribution has shifted from content creators to content distributors — social media and platform companies — that select which articles and messages we see first and next. Their selection algorithms, known as *recommenders*, are largely tuned to keep our attention, to engage versus inform us based on our past online behavior. Platforms have a duty to perform well financially, and it is by capturing our attention that they maximize profits from advertisers. Yellow journalism is by no means a new phenomenon, but the scale of platforms and the accuracy of targeting achieved through AI algorithms has greatly increased the consequences of inaccurate, unverifiable, or contextually misleading information.

Recommenders have major flaws: they rarely present content in chronological order; they reduce ideas down to easily consumable images, videos, and snippets of text; they prioritize highly

---

[1] The term *information integrity* refers to the accuracy and reliability of information. It has been adopted in place of misinformation and disinformation as it more broadly addresses the state of information, and the consequences, as opposed to the intent of its distribution. *Misinformation* is usually defined as inaccurate or deceptive information irrespective of intent; *disinformation* is narrower, comprising malicious intent. Importantly, in the context of online social networks, the term *integrity* has been adopted as a desired goal, representing an ecosystem in which users are protected from a range of harms such as misinformation, hate speech, illegal markets, and exploitation. See *Journalism, Fake News & Disinformation. UNESCO, 2018. https://bit.ly/2MuELY5 and Preserving Integrity in Online Social Networks. arXiv:2009.10311. https://arxiv.org/abs/2009.10311.*

[2] Google, Facebook, Twitter, Apple, Amazon, and other platforms control the spread of ads, articles, videos, photos, updates, and other forms of messages between billions of commercial customers and individual users.

charged, outrageous, and polarizing messages over well-researched objective information; and they prioritize content that reinforces existing beliefs. Together these flaws increase misleading narratives, "filter bubbles"[3], and social polarization.

In this ecosystem, narratives gain authority based on the appearance and style of presentation and perceived acceptance by social cliques, rather than based on verifiable sourcing from one of the few journalistically rigorous, authoritative news sources of the past. Furthermore, any narrative can be amplified with ad spend. Through recommenders, every sponsor of a narrative has, literally at their fingertips, a powerful amplifier capable of greater influence on a larger crowd per dollar per second than any prior medium in history.[4] This advertising ecosystem becomes increasingly problematic as disinformation has been found to be more "engaging" than accurate information, thus driving a profitable industry of both willing and unwitting sponsors.[5]

The offline consequences in this new context are alarming.

Disinformation mobilized an attack on the US Capitol. Mis- and disinformation led to the January 6th attack on the US capitol in Washington DC, in part encouraged by a QAnon-supported conspiracy that former President Donald Trump's re-election was "stolen." Just a month before the attack the broad-scale conspiracy of a stolen election was either accepted as true or not rejected as false, by more than half of Americans.[6] In 2019, QAnon's connection to violence prompted the FBI to classify it as a domestic terrorism threat.

Mis- and disinformation have deterred our ability to protect public health, the climate, and democracy. It has hampered our ability to manage the COVID pandemic. Thirty percent of Americans believe at least one COVID-19 conspiracy,[7] and mis- and disinformation are the largest drivers of vaccine hesitancy.[8] Progress on climate change faces perpetual threats by special interest groups using inaccurate information to push an agenda.[9] The loss of information

---

[3] The term *filter bubble* was coined by Eli Parser in 2011 to describe algorithmically-induced intellectual isolation from viewing content that simply reinforces one's perspectives. See *The Filter Bubble: How the New Personalized Web Is Changing What We Read and How We Think* by Eli Parser.

[4] Doty, David. "It's All About Pricing: Digital Is Winning Simply Because It's A Cheaper Way For Advertisers To Reach Consumers: A 101 Course." Forbes. Forbes Magazine, October 29, 2019. https://www.forbes.com/sites/daviddoty/2019/10/29/its-all-about-pricing-digital-is-winning-simply-because-its-a-cheaper-way-for-advertisers-to-reach-consumers-a-101-course/?sh=49176ee73275.

[5] The Global Disinformation Index's Report. "GDI Primer: The U.S. (Dis)Information Ecosystem," October 2020.

[6] See, for example, this Ipsos poll of Dec 30, 2020

[7] Uscinski, Joseph E., Adam M. Enders, Casey Klofstad, Michelle Seelig, John Funchion, Caleb Everett, Stefan Wuchty, Kamal Premaratne, and Manohar Murthi. "Why Do People BELIEVE COVID-19 Conspiracy Theories?: HKS Misinformation Review." Misinformation Review, January 26, 2021. https://misinforeview.hks.harvard.edu/article/why-do-people-believe-covid-19-conspiracy-theories/.

[8] Sgaier, Sema K. "Meet the Four Kinds of People Holding Us Back from FULL VACCINATION." The New York Times. The New York Times, May 18, 2021. https://www.nytimes.com/interactive/2021/05/18/opinion/covid-19-vaccine-hesitancy.html.

[9] Zakrzewski, Cat. "The Technology 202: How Social Media Helped FUEL False Claims about Texas Power Outages." The Washington Post. WP Company, February 23, 2021. https://www.washingtonpost.com/politics/2021/02/23/technology-202-how-social-media-helped-fuel-false-claims-about-texas-power-outages/.

integrity in our social platforms has also been used to undermine our democracy directly. During the final days of the 2020 election, disinformation campaigns aiming to discourage voting were targeted specifically at Black and Latino voters.[10] Following the election, disinformation campaigns supporting a narrative of voter fraud have driven the passing of the strictest voter suppression laws in decades around the country.[11]

Network Effects vs Protected Speech

The flows of mis- and disinformation through social networks are amplified by network effects—reaching exponentially larger crowds at each step. Once a conspiracy theory takes root among a group of spreaders, it leads to the formation of a social movement that becomes nearly impossible to curb. A malicious narrative tends to have one or a handful of originators, and then a small number of influential spreaders, each with thousands of followers. The digital platform firms play a critical role in any effort to detect, trace, and eventually mitigate its spread. This would suggest a simple solution: trace inaccurate information to its sources and cut the problem off at its source. But in the US, we must balance the duty of care to minimize harm with the right to dissenting speech.

In the massively amplified, ultra-connected information ecosystem, any party spreading mis- and disinformation has unprecedented influence. Whereas cult members on street corners might recruit one new member over the course of a day by engaging passers-by, a convincing message in a video can act as a dragnet, reaching millions through a small number of influencers online.

To detect and throttle harmful content while protecting the free expression of dissent requires understanding the actors operating in the information ecosystem. **Accidental Spreaders** are actors unaware that they are spreading mis- or disinformation. Accidental spreaders will also differ as some may be less willing than others to accept the inaccuracies of the information they shared. Assessing the degree of influence that mis- and disinformation has on individuals is critical to reintegrating people into accurate information pipelines. **Malicious Spreaders** are actors intentionally spreading disinformation and gaining from its spread. Malicious spreaders can be individuals, for example, alternative health entrepreneurs profiting from spreading COVID vaccine falsities,[12] or more systemic spreaders such as governments and political campaigns looking to undermine elections.

The government's duty of care to protect its people from harm is proportional to the influence of an actor in spreading narratives. One technical remedy in the case of digital platform companies is, rather than risking the complete loss of free expression, to reduce its reach by throttling its

[10] Bond, Shannon. "Black and Latino Voters Flooded with Disinformation IN ELECTION'S Final Days." NPR. NPR, October 30, 2020. https://www.npr.org/2020/10/30/929248146/black-and-latino-voters-flooded-with-disinformation-in-elections-final-days.

[11] See, The Brennan Center for Justice's Voting Laws Roundup: March 2021.

[12] Bond, Shannon. "Just 12 People Are Behind MOST Vaccine Hoaxes on Social Media, Research Shows." NPR. NPR, May 14, 2021. https://www.npr.org/2021/05/13/996570855/disinformation-dozen-test-facebooks-twitters-ability-to-curb-vaccine-hoaxes.

spread down to a level commensurate with the framers' intent, from, say, one million recipients per hour to a hundred.

Technical Solutions

Policies that thread the needle of protected speech are irrelevant without the practical means to govern the new digital information ecosystem. We see a set of technical tools and methods,[13] many of which are in use today, as key to a plan for restoring information integrity, which starts with recognizing offline harms and ends with minimizing the spread of inaccurate, unverifiable, and contextually misleading information.

Track offline harms from online information spread. To justify any interference in the exchange of information, demonstrating the harm of not interfering is critical. This is why the first class of technical solutions to mis- and disinformation is technology to measure and track harms. This is an active area of research, with an imperfect record due to conflicts between social media companies and researchers. The Cambridge Analytica scandal, in which the firm worked with a private researcher to harvest the data of millions of Facebook users by taking advantage of a privacy loophole in Facebook's policy,[14] has caused platform companies to eschew sharing data with researchers. Facebook recently disabled data access for researchers working with the NYU Ad Observatory to study political ad misinformation, claiming a violation of its terms.[15] Despite ongoing tensions, the means for tracking the offline harms of the digital information ecosystem exist and continue to be developed by researchers and digital platforms themselves. While causal links between online mis- and disinformation and offline harm are difficult to prove, there are useful, available surrogate measures based on analyzing data available to the digital platform companies. Tracking correlation, if not causation, is one of the strengths of today's AI algorithms: for example, one can track, among criminal cases, the number of those convicted who had joined online extremist groups; among COVID-19 cases and deaths, the number who had eschewed vaccination and were exposed to pandemic misinformation. These correlations are significant as they inform testable hypotheses which, if proven, can shape public and private policies. House Bill 8636, the Protecting Americans from Dangerous Algorithms Act, proposes an amendment to Section 230 most likely to pass, which makes digital platform companies liable for online content that leads to offline terrorist activity.[16] The technical means to track offline

---

[13] For a comprehensive look at misinformation interventions, see: Saltz, E., & Leibowicz, C. (2021, June 14). *Fact-checks, info hubs, and shadow-bans: A landscape review of misinformation interventions*. Partnership on AI. Retrieved October 1, 2021, from https://partnershiponai.org/intervention-inventory/.

[14] Rosenberg, Matthew, Nicholas Confessore, and Carole Cadwalladr. "How Trump Consultants Exploited the Facebook Data of Millions." The New York Times, March 17, 2018. https://www.nytimes.com/2018/03/17/us/politics/cambridge-analytica-trump-campaign.html.

[15] Ortutay, Barbara. "Facebook Shuts Down NYU Academics' Research on Ads, Citing 'Data Scraping'." Los Angeles Times, August 5, 2021. https://www.latimes.com/world-nation/story/2021-08-05/facebook-shuts-down-nyu-academics-research-on-ads-citing-data-scraping.

[16] House of Representatives Bill, H.R.8636 - Protecting Americans from Dangerous Algorithms Act.

harms from online information will be key to enforcement of mis- and disinformation-related regulation.

Enable responsible algorithm design and review. The internal design of digital platforms does not encourage information integrity.[17] Social platforms, such as Facebook and Instagram, prioritize "engagement" measured through user interactions (i.e.likes, comments, and shares on posts) furthering the promotion of sensational content.[18] Conspiracies and extreme content drive ad revenue by virtue of being more "engaging" than authoritative posts.[19] Recommenders are among the most impactful AI algorithms[20] not currently subject to external risk assessment, design standards, or 3rd-party review. This is a key area of technical opportunity to address the information ecosystem and uphold integrity standards.

Increase the transparency of content by adding labels. Today's AI offers sophisticated models that assess image and text together to classify multimodal content, and the digital platform companies rely on them for internal labeling.[21] Although we do not expect full automation, technical solutions will serve a key role in clustering content into categories for human labeling. For example, a system can sort millions of message posts and videos into a coarse-to-fine hierarchy of subject areas that humans can then use to judge the nature of the posts. Judges can similarly use technology to speed the process of categorizing the sponsor and sources (which may differ); information about how much was paid; and determine characterizations such as "likely to incite violence" and "known to conflict with the advice of the surgeon general"; and ultimately suggest verifiable alternatives from credible sources. While labeling mis- and disinformation alone is not sufficient for discontinuing its spread,[22] it can deter unintentional dissemination. Such labeling applies the norms of analog news,[23] which differentiates objective information from opinion-based pieces, and applies the same standards to digitally disseminated information.

---

[17] Hao, Karen. "How Facebook Got Addicted to Spreading Misinformation." MIT Technology Review. MIT Technology Review, March 11, 2021. https://www.technologyreview.com/2021/03/11/1020600/facebook-responsible-ai-misinformation/.

[18] Frier, Sarah, and Sarah Kopit. "Facebook Built the Perfect Platform for Covid Vaccine Conspiracies." Bloomberg.com. Bloomberg, April 1, 2021. https://www.bloomberg.com/news/features/2021-04-01/covid-vaccine-and-fertility-facebook-s-platform-is-letting-fake-news-go-viral.

[19] Rosalsky, Greg. "Are Conspiracy Theories Good for Facebook?" NPR. NPR, August 4, 2020. https://www.npr.org/sections/money/2020/08/04/898596655/are-conspiracy-theories-good-for-facebook.

[20] Stray, Jonathan. "Beyond Engagement: Aligning Algorithmic Recommendations with Prosocial Goals." The Partnership on AI, January 21, 2021. https://www.partnershiponai.org/beyond-engagement-aligning-algorithmic-recommendations-with-prosocial-goals/.

[21] Facebook Research. "Hateful Memes Challenge and Dataset." Facebook AI, May 12, 2020. https://ai.facebook.com/blog/hateful-memes-challenge-and-data-set/. ;
Facebook. "Here's How We're Using Ai to Help Detect Misinformation." Facebook AI. Accessed August 12, 2021. https://ai.facebook.com/blog/heres-how-were-using-ai-to-help-detect-misinformation/.

[22] Saltz, Emily, and Claire Leibowicz. "Labeling Misinformation Isn't Enough. Here's What Platforms Need to Do next." The Partnership on AI, March 11, 2021. https://www.partnershiponai.org/labeling-misinformation-isnt-enough/.

[23] Karen Kornbluh, Ellen P. Goodman. "Opinion | Three Steps to Help Treat America's Debilitating Information Disorder." The Washington Post. WP Company, January 13, 2021. https://www.washingtonpost.com/opinions/2021/01/13/three-steps-help-treat-americas-debilitating-information-disorder/.

Throttle algorithmic recommendations to minimize the spread of harmful information.[24] Another key technical means to control disinformation is to reduce the speed of spread based on the potential for harm. Platform companies can use various methods to measure the reliability of an account based on posting history, such as implementing integrity scores based on posting history (i.e., the more "deep fakes" a profile posts, the lower the profile's integrity score). The scores can serve as one of several measures to throttle the algorithmic reach of a profile, with posts from low-score profiles being less widely disseminated. Second, the spread of inaccurate *posts,* regardless of profile*,* can be throttled to reduce views. A similar notion — the "circuit breaker" — was introduced by Kornbluh and Goodman (2020) with Facebook and Twitter adopting variations of post throttling.[25] Throttling content based on integrity scores augments prior recommendations by expanding content moderation beyond ad hoc removal, towards actively promoting accurate, authoritative content and users.

Establish a track and trace system to find the originating source of problematic content and alert users who were exposed to it. Tracking the sources of mis- and disinformation is critical for informing account integrity scores, which determine account reach; understanding the ways in which types of information spread, and; creating and enforcing liability. Technical methods such as "clustering," grouping posts together based on similarity using language analysis models, for instance, can aid in implementing a track and trace system. Researchers at MIT have already developed the Reconnaissance of Influence Operations (RIO) program, which automatically detects and analyzes social media accounts spreading disinformation.[26] Additionally, alerting users who were exposed to and interacted with unverifiable information is critical to slowing the unintended spread of misinformation and increasing information awareness. This has been done to a small, decentralized degree by Twitter, which notified users who had interacted with tweets from the Internet Research Agency (IRA) known for pushing Russian propaganda, and more recently, by Facebook, which is testing alerts to users exposed to extremist content.[27]

Conclusion

[24] Stray, J. Aligning AI Optimization to Community Well-Being. *Int. Journal of Com. WB* 3, 443–463 (2020). https://doi.org/10.1007/s42413-020-00086-3

[25] Karen Kornbluh and Ellen P. Goodman. "Safeguarding Digital Democracy," German Marshall Fund of the United States, March 24, 2020.

[26] McGovern, Anne. "Artificial Intelligence System Could Help Counter the Spread of Disinformation." MIT News | Massachusetts Institute of Technology, May 27, 2021. https://news.mit.edu/2021/artificial-intelligence-system-could-help-counter-spread-disinformation-0527.

[27] Rosenberg, Eli. "Twitter to Tell 677,000 Users They Were Had by the Russians. Some Signs Show the Problem Continues." The Washington Post. WP Company, April 8, 2019. https://www.washingtonpost.com/news/the-switch/wp/2018/01/19/twitter-to-tell-677000-users-they-were-had-by-the-russians-some-signs-show-the-problem-continues/. ;
BBC Tech. "Facebook Tests Extremist Content Warning Messages." BBC News. BBC, July 2, 2021. https://www.bbc.com/news/technology-57697779.

Addressing the growing societal harms of the online information ecosystem requires a comprehensive plan with corresponding practical technology solutions. Such a plan requires organized action across government, platforms, and civil society. The use of technology is a key factor in recognizing offline harms and controlling the dissemination of mis- and disinformation. This proposal offers a set of technical solutions that can assist in an urgently needed strategy for restoring information integrity.

APPENDIX: Progress in Regulation to Address Disinformation

The market in online influence is so much more efficient and effective than oratory, print, radio, and television that its potential harms cannot sufficiently be addressed by minor amendments to existing US regulation. Relevant US law has not changed in 20 years, while technological progress has transformed industries and social norms. These factors drive a pressing need for government action.

Across Europe

The European Commission has proposed ground-breaking AI and platform regulation that together begin to address information integrity. In the last two years, the EC has launched the European Democracy Action Plan (EDAP) as well as proposed the Digital Services Act (DSA), Guidelines for Strengthening the Code of Practice on Disinformation, and a framework for AI regulation. Together, they represent an ambitious approach to regulating the digital ecosystem, including combatting disinformation by prohibiting harmful actions and establishing transparency, reporting, and accountability mechanisms. For instance, the Guidance for Strengthening the Code of Practice on Disinformation goes as far as to suggest that platforms make recommendation algorithms not only transparent regarding their content prioritization factors but user-customizable.

While these regulations include elements of a robust package to combat disinformation, they still fall short of the sweeping approach necessary. Within the DSA, for instance, only very large online platforms (VLOPs) are subject to most disinformation-based regulations, such as the obligations around recommender system transparency.[28] This will undoubtedly create barriers to upholding information integrity as disinformation is a cross-platform phenomenon, not limited to VLOPs.[29] Nonetheless, Europe's proposed regulations serve as a useful starting point for upholding information integrity within the United States; and critically, doing so at a Federal level, as a set of united states.

In the United States

Section 230

Section 230 of the Communications Decency Act shields online services from liability for the content they distribute.[30] Credited with enabling Big Tech to become a trillion-dollar sector, the

[28] Very Large Online Platforms are those that reach more than 10% of 450 million consumers in Europe.

[29] For a detailed review of the DSA's limitations around combatting disinformation, as well as suggestions for improving the regulation, see How the Digital Services Act (DSA) Can Tackle Disinformation.

[30] Kosseff, Jeff. *The Twenty-Six Words That Created the Internet*. Ithaca, NY: Cornell University Press, 2019.

law allows, for example, Twitter and Facebook to support and amplify the viral spread of disinformation, causing offline harms such as unnecessary COVID deaths and the January 6th insurrection without liability. Since the 2016 election, many amendments have been proposed to remedy the perceived problems with Section 230, despite fierce pushback from Big Tech.[31]

Notably, among the 26 proposed section 230 bills introduced in the 116th congress, just one reintroduced in the 117th is close to becoming US law.[32] The "Protecting Americans from Dangerous Algorithms Act" lifts Big Tech's liability shield when offline violence results from their algorithmic promotion of harmful, radicalizing content. The bill simply moves the liability for speech that leads to terrorist acts from resting solely with the originator to resting with both the originator and social media firms acting as amplifiers via opaque promotion algorithms. However, the concentration of proposals on Section 230 has led many expert regulators to doubt its viability as a legislative solution.

State Level Approaches

Individual states have also introduced relevant, albeit perfunctory, legislation. The California State Assembly, for instance, recently passed AB 587 which requires social media platforms to display their terms of service in a specified manner to highlight the platforms' policies aimed at countering false information, harassment, hate speech, extremism, and protecting users from foreign interference.[33] Additionally, New York lawmakers recently proposed bills S.4511, S.4512, and S.4531 which require social media networks to provide and maintain mechanisms for reporting hateful conduct, vaccine disinformation, and election disinformation respectively.[34]

While liability is the regulatory lever that has received the most attention in Congress, broad oversight is a critical but underexplored solution. There have been external proposals for a new agency that would establish requirements of transparency and due process; measure harms from algorithms and content; provide timely and effective regulatory review, and; enforce accountability mechanisms.[35] We echo calls for the establishment of a permanent agency aimed at ensuring National Information Integrity.

Thus far, the proposed US legislation ignores the duty of care necessary due to scale and automation. The scale of connected devices combined with the automation that enables

---

[31] See, Facebook's Whitepaper Charting a Way Forward: Online Content Regulation.

[32] Congressional Research Service, Rep. *Social Media: Misinformation and Content Moderation Issues for Congress*, 2021.

[33] California State, Assembly Bill 587

[34] New York State bills S.4511, S.4512, and S.4531

[35] Tutt, Andrew. "An FDA for Algorithms." *SSRN Electronic Journal*, March 15, 2016. https://doi.org/10.2139/ssrn.2747994. ; Whelan, Moira, and Vera Zakem. "America Needs a New Way to Combat Disinformation Now." Foreign Policy, January 22, 2021. https://foreignpolicy.com/2021/01/22/united-states-capitol-siege-disinformation-commission/.

mass-custom influence creates a legal and ethical duty of care from which content distributors currently enjoy immunity.