Measuring the news and its impact on democracy

Duncan J. Watts[a,b,c,1], David M. Rothschild[d], and Markus Mobius[e]

[a]Department of Computer and Information Science, University of Pennsylvania, Philadelphia, PA 19104; [b]The Annenberg School of Communication, University of Pennsylvania, Philadelphia, PA 19104; [c]Operations, Information, and Decisions Department, University of Pennsylvania, Philadelphia, PA 19104; [d]Microsoft Research, New York, NY 10012; and [e]Microsoft Research, Cambridge, MA 02142

Edited by Dietram A. Scheufele, University of Wisconsin–Madison, Madison, WI, and accepted by Editorial Board Member Susan T. Fiske February 21, 2021 (received for review November 8, 2019)

Since the 2016 US presidential election, the deliberate spread of misinformation online, and on social media in particular, has generated extraordinary concern, in large part because of its potential effects on public opinion, political polarization, and ultimately democratic decision making. Recently, however, a handful of papers have argued that both the prevalence and consumption of "fake news" per se is extremely low compared with other types of news and news-relevant content. Although neither prevalence nor consumption is a direct measure of influence, this work suggests that proper understanding of misinformation and its effects requires a much broader view of the problem, encompassing biased and misleading—but not necessarily factually incorrect—information that is routinely produced or amplified by mainstream news organizations. In this paper, we propose an ambitious collective research agenda to measure the origins, nature, and prevalence of misinformation, broadly construed, as well as its impact on democracy. We also sketch out some illustrative examples of completed, ongoing, or planned research projects that contribute to this agenda.

misinformation | media | democracy

It is hard to overstate the breadth and intensity of interest directed over the past 2 y at the issue of false or misleading information (also known as "fake news") circulating on the web in general and on social media platforms such as Facebook and Twitter in particular (1–13). According to Google Scholar, since January 2017, more than 5,000 English language publications with "fake news" in the title have appeared in academic journals spanning economics, political science, computer and information science, communications, law, and journalism. To put this number in perspective, fewer than 100 such publications appeared in all the years leading up to the end of 2016, while fewer than 600 publications have appeared since 2017 containing "television news" or "TV news."

The origin of this extraordinary surge in interest in a previously sleepy topic was of course the 2016 US presidential election, which, along with other events that year such as Brexit, raised widespread concerns about a possible rise of populist/nationalist political movements, increasing political polarization, and decreasing public trust in the media. Early reporting by journalists (14) quickly focused attention on fake news circulating on social media sites during the election campaign. The philanthropic and scientific communities then responded with dozens of conferences and thousands of papers studying various elements of fake news. Reinforced by continued mainstream media attention and increasing congressional scrutiny of technology companies, the conjecture that the deliberate spread of online misinformation poses an urgent threat to democracy subsequently hardened into conventional wisdom (13, 15).

In the face of this dominant narrative, a handful of authors (1, 10, 11, 16–18) have suggested that fake news is less prevalent than breathless references to "tsunamis" or "epidemics" would imply. In an early contribution, Allcott and Gentzkow (1) estimated that "the average US adult read and remembered on the order of one or perhaps several fake news articles during the election period, with higher exposure to pro-Trump articles than pro-Clinton articles." In turn, they estimated that "if one fake news article were about as persuasive as one TV campaign ad, the fake news in our database would have changed vote shares by an amount on the order of hundredths of a percentage point," roughly two orders of magnitude less than needed to influence the election outcome. Subsequent studies have found similarly low prevalence levels for fake news relative to mainstream news on Twitter (10) and Facebook (11). Finally, our own survey of the media consumption landscape, based on a nationally representative sample of TV, desktop, and mobile media consumption (18), found three main results that undercut the conventional wisdom regarding fake news and also the dominance of online sources of news in general:

1) News consumption is a relatively small fraction of overall media consumption. Of the more than 7.5 h per day that Americans spend, on average, watching television of consuming content on their desktop computers or mobile devices, only about 14% is dedicated to news ("news" was defined as appearing on one of more than 400 news-relevant programs [e.g., *CBS Evening News*] and more than 800 websites [e.g., http://www.nytimes.com/], while "consumption" was measured in terms of minutes per person per day watching television or browsing online; see ref. 18 for details).
2) Online news consumption is a small fraction of overall news consumption, which is dominated by TV by a factor of five to one. Even 18 to 24 y olds consume almost twice as much TV news than online news. In striking contrast with the research literature's overwhelming emphasis on online sources of news, we estimate that three in four Americans spend less than 30 s a day reading news online, while almost half consume no online news whatsoever.
3) Fake news is a tiny portion of Americans' information diets. Using our most inclusive definition, less than 1% of regular news consumption and less than 1/10th of 1% of overall media consumption could be considered fake. Even the heaviest consumers of fake news (the 55+ age group) consume less than 1 min of fake news per day on average, compared with

This paper results from the Arthur M. Sackler Colloquium of the National Academy of Sciences, "Advancing the Science and Practice of Science Communication: Misinformation About Science in the Public Sphere," held April 3–4, 2019, at the Arnold and Mabel Beckman Center of the National Academies of Sciences and Engineering in Irvine, CA. NAS colloquia began in 1991 and have been published in PNAS since 1995. From February 2001 through May 2019, colloquia were supported by a generous gift from The Dame Jillian and Dr. Arthur M. Sackler Foundation for the Arts, Sciences, & Humanities, in memory of Dame Sackler's husband, Arthur M. Sackler. The complete program and video recordings of most presentations are available on the NAS website at http://www.nasonline.org/misinformation_about_science.

Author contributions: D.J.W., D.M.R., and M.M. designed research; D.J.W., D.M.R., and M.M. performed research; and D.J.W. and D.M.R. wrote the paper.

The authors declare no competing interest.

This article is a PNAS Direct Submission. D.A.S. is a guest editor invited by the Editorial Board.

[1]To whom correspondence may be addressed. Email: djwatts@seas.upenn.edu.

Published April 9, 2021.

106 min of regular news (94 of them on TV) and over 500 min of total media consumption.

As has been argued elsewhere (13), these results on their own do not conclusively demonstrate that fake news does not have meaningful effects on public opinion, political polarization, and trust in institutions. It is possible, for example, that even extremely low rates of exposure to fake or misleading news could have outsized effects, at least on some people, or that equivalent amounts of online and television news consumption have different impacts. Nonetheless, these results do strongly suggest that research on the origins, nature, prevalence, and consequences of misinformation should take a much broader view of the topic than outright false information disseminated on social media or even online (16). In particular, there are at least three reasons for taking such a broader view.

First, while it is possible that exposure to fake news has more impact than an equivalent amount of exposure to real news, or that online news has more impact than television news, it is equally possible that the opposite is true. For example, recent work has found that subjects rate mainstream publications as more trustworthy than fake or highly partisan sites irrespective of their own partisanship (12), and that deliberation reduces belief in false headlines but not in true ones, again irrespective of partisan alignment (19). Likewise, while television consumption can be dismissed as more "passive" than reading, direct comparisons between television and online news and advertising consistently find better recall of televised content (20–22) especially for low-involvement consumers (23). Ultimately, questions of impact are empirical questions and answering those questions will require making comparisons between different types of content and different modes of production.

Second, fake news sites are not the only sources of false information: The mainstream media can also promulgate falsehoods simply by reporting on them (24). In the lead-up to the 2003 Iraq War, for example, a large majority of media organizations uncritically repeated the administration's false claim that they possessed unequivocal evidence that Saddam Hussein possessed weapons of mass destruction (25, 26). In August 2009, when Sarah Palin wrote in a blog post that the Affordable Care Act would create "death panels," the claim was repeated in over 700 mainstream news articles even after it was debunked by a variety of fact-checking organizations (27, 28). More recently, an analysis of Russian disinformation efforts during the 2016 presidential election concluded that these efforts likewise succeeded in reaching the public largely via the credulous reporting of mainstream media outlets (29). Although the motivations and mechanisms driving misinformation in mainstream media differ from sites that intentionally promote falsehoods, the effects may be many times greater; thus, a proper accounting of the prevalence of false information requires a broad consideration of potential sources.

Third, misinformation is a much broader phenomenon than outright falsehoods. There are many ways to lead a reader (or viewer) to reach a false or unsupported conclusion that do not require saying anything that is unambiguously false (30). Presenting partial or biased data, quoting sources selectively, omitting alternative explanations, improperly equating unequal arguments, conflating correlation with causation, using loaded language, insinuating a claim without actually making it (e.g., by quoting someone else making it), strategically ordering the presentation of facts, and even simply changing the headline can all manipulate the reader's (or viewer's) impression without their awareness. These practices are pervasive in mainstream professional journalism (see, e.g., ref. 31) and are not restricted to political topics, although that is often the focus of research on media bias (32, 33). Inaccurate and misleading coverage is also pervasive in other areas of journalism (34), including important domains for public opinion and democracy such as health (35, 36), science (37), and business (38).

For all three reasons, studies of the prevalence of misinformation and its impact on democratic decision making must embrace a much broader conception of the problem that includes biased and potentially misleading information that is embedded in mainstream news content across all major modes of production (24, 29). Unfortunately, research of this scope and scale is hindered by three interrelated but distinct obstacles. First, research on misinformation and its effects is currently dependent on datasets that are idiosyncratic, one-off, and often small in scale, rendering comparisons across different modes of media consumption, different sample populations, and different time periods difficult to make. Second, much of the relevant data are hard to collect, either because they are scattered across thousands of locations in different formats, or are controlled by private companies (Google, Facebook, Twitter, Microsoft, media companies, etc.) who face large disincentives and limited upsides to sharing data with academic researchers (39). Third, the relevant academic research is scattered across several disciplines (e.g., economics, marketing, political science, communications, psychology, sociology, computer science, and network science), each with its own set of theoretical frameworks, accepted methodologies, and publishing venues. Collating and reconciling results across these disciplinary boundaries is difficult and often leads to contradictory or incoherent conclusions (40).

Addressing these shortcomings in existing data and research practices will require a major effort to coordinate scientific communities, data resources, and academic–industry collaborations. Although we are not the first to call for such an effort (see, e.g., refs. 4, 5, and 39), our proposal differs from previous instantiations in that it is explicitly focused on the need for shared research infrastructure as well as the opportunities for collaboration and partnership that such an infrastructure may create. In the next section, we describe our proposal at a high level, breaking it into four distinct but mutually reinforcing objectives. We then illustrate the potential of our proposed approach with a selection of in-progress and planned research projects, as well as some examples of public outreach and engagement projects that we believe will enhance the research.

Toward a Comprehensive Misinformation Research Agenda

The objective of a comprehensive research agenda to study the origins, nature, and consequences of misinformation on democracy in turn entails assembling four subsidiary components:

1) A large-scale data infrastructure for studying the production, distribution, consumption, and absorption of news over time and across the entire information ecosystem (including the web, television, radio, and other modes of production).
2) A "mass collaboration" model that leverages the shared infrastructure to advance replicable, cumulative, and ultimately useful science.
3) A program for communicating the insights generated by the research to stakeholders outside of the research community (e.g., journalists, policymakers, industry leaders, the public).
4) A network of academic–industry partnerships around data and solutions.

Objective 1: Building a Large-Scale Data Infrastructure for Studying News Production, Distribution, Consumption, and Absorption. A primary requirement for comprehensive research agenda around misinformation is a shared, open infrastructure for collecting data and running experiments at scale for diverse populations over long timescales. Such an infrastructure would facilitate results that generalize better than prior work and can be more easily implemented in practice. Moreover, the infrastructure

would be open, meaning that it would be made available to the research community while also addressing issues of data security, individual privacy, and intellectual property. To illustrate the scale and scope of the proposed infrastructure, Fig. 1 shows a schematic of the information ecosystem, which is represented in four "layers": 1) production, 2) consumption and distribution, 3) absorption and understanding, and 4) action and engagement. Each layer corresponds to a different stage of the process by which information about events and issues affecting a democracy ultimately impacts public opinion, understanding, and civic engagement. Each layer also corresponds to different types of data that derive from distinct sources, typically in different formats and sampled in different ways.

Production (web, TV, radio). What information is produced, either by online publishers or by TV or radio broadcasters, that could potentially inform and/or influence public opinion? The web alone comprises many thousands of news sources, ranging from large and comprehensive (e.g., *The New York Times*, *The Wall Street Journal*) to small and niche, from neutral to partisan, and including original news publishers as well as aggregators and distributors. As noted earlier, publishers can bias the news they produce in several ways, including selection (what they choose to cover vs. ignore), emphasis (how prominently a given story is featured and for how long), slant (how headlines are written, the tone of the article, the relative emphasis of different facts), and finally outright deception (fake news, propaganda, etc.). To obtain a comprehensive, longitudinal view of information production, the research community requires a continuously updated catalog of information sources relevant to contemporary issues and political discourse.

Several media databases already exist (e.g., Media Cloud, Event Registry, GDELT, Internet Archive's TV news archive, Newsbank). However, they are not designed to directly support the range of queries that are the focus of many research questions; thus, results typically require substantial investment in postprocessing. In addition, they do not exhibit the kind of methodological transparency that is required for academic research (41) and/or they do not have the comprehensiveness across the necessary range of site and modes. To illustrate the problem, simple keyword searches (e.g., "Hillary Clinton emails") on unpreprocessed corpora of articles will return many irrelevant articles (i.e., those that contain the keywords but are not about the topic) and will also miss many relevant articles (i.e., those that are on the topic but do not use the exact keywords). Moreover, the results contain no information about features such as partisanship or sentiment that must then be appended by the researcher. Keyword-based search results, in other words, are largely uninformative without a large amount of supplemental data cleaning and analysis. Because this work is typically done in a one-off, nonreplicable manner, simply collecting and storing vast amounts of news data does not on its own do much to accelerate the research process. A central objective for any collective research effort, therefore, is to build data processing pipelines and systems on top of the raw data that make them easily queryable by researchers and journalists alike. Included in this objective is also the capability for independent researchers to develop and contribute new modes of querying (e.g., abstracting away from specific stories to broader themes or narratives) as well as new methods for generating relevant metadata (e.g., stance, sentiment, partisan bias, etc.).

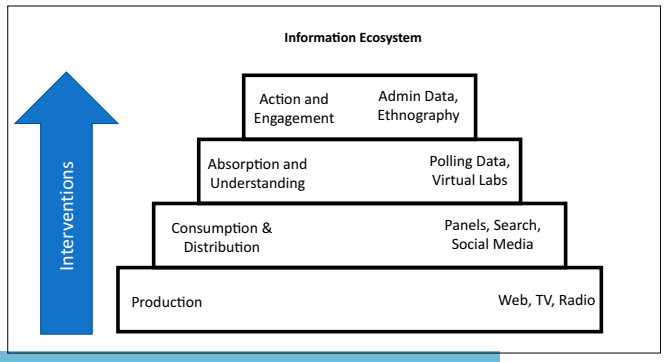

Fig. 1. Schematic representation of the information ecosystem.

Consumption and distribution (desktop and mobile panels). Much of the information that is produced receives little attention, while some stories resonate with millions. Even comprehensive and well-annotated data on news production, therefore, do not on their own tell us how that information is or is not reaching consumers, let alone how different types of information reaches different types of consumers. Are there groups of people who watch MSNBC in the morning, surf mainstream news during the day, and watch Fox News at night? Do Breitbart and Daily Kos readers also get mainstream news on TV or the web? One potential direction for research on media consumption is to leverage commercial panel providers such as Nielsen, ComScore, Pew Research, and YouGov. Although valuable (see, e.g., next section), these "off-the-shelf" solutions also exhibit some important limitations. In particular, desktop-only panels increasingly suffer from coverage gaps in part because they do not capture mobile activity, and in part because an increasing amount of web traffic is contained in "walled gardens" such as Facebook within which user activity is visible only to the platform. Ultimately, therefore, it will be necessary to develop new data sources. For example, a dedicated mobile panel would greatly facilitate the measurement of information consumption across social and conventional media, as well as enable linkage to other behaviors of potential interest. In addition, certain modes of consumption—in particular social media (e.g., Facebook, Twitter, Reddit), but also email, messaging services (e.g., WhatsApp)—are also mechanisms for distribution. A proper understanding of consumption, therefore, will also require data on information distribution.

Absorption and understanding (polls, virtual labs). Just as the publication of a particular piece of information does not guarantee that anyone will see it, so is exposure to information no guarantee of awareness, understanding, or agreement about its meaning (4, 42). Exposure to disconfirming information may reduce polarization, increase it, or have no effect depending on other factors (43, 44). Understanding how consumption translates into knowledge and/or beliefs is therefore critical to designing and evaluating possible interventions. Building off of recent advances in nonprobability polling techniques (45, 46), one could conduct regular panel surveys to probe public knowledge and explore the baselines and shifts in knowledge and attitudes. Polling of this sort could yield indices of facts and sentiment from the general population that could be correlated with media consumption on various issues and, ultimately, civic participation. Understanding of opinion change, influence, and deliberation would also be accelerated via experiments conducted in online "virtual labs" (47).

Action and engagement (admin data, ethnography). In addition to being an end in itself, knowledge is also important to democracy inasmuch as it translates into political action: voting, community organizing, engagement with legislators, political speech, and protest. An important goal for any comprehensive research agenda is therefore to understand the link between the production, consumption, and absorption of information on the one hand, and action on the other hand. Because "political action" is a multidimensional concept, however, quantifying action is challenging, at a minimum requiring diverse administrative datasets (e.g., voter records, campaign contributions, volunteering, protesting, search, activity on social media, etc.), but also

https://doi.org/10.1073/pnas.1912443118

survey and ethnographic data to elucidate levels of engagement in the political process, broadly construed (48). Alternatively, or in order to get repeated actions or more coverage, researchers could leverage proxies for engagement such as search queries as a proxy for intent (49) or lightweight user actions (following, retweeting, liking, commenting, etc.) as a proxy for interest (50).

Objective 2: Build a "Mass Collaboration" Model to Advance Replicable, Cumulative, and Useful Knowledge. Maximizing the value of the data infrastructure just described will also require a "mass collaboration" model in which many researchers leverage the same data assets (51). Mass collaboration models based on shared infrastructure have an established track record in the physical sciences (e.g., the Sloan Digital Sky Survey, the Large Hadron Collider, the Laser Interferometer Gravitational-Wave Observatory) and also in biology (e.g., the Human Genome Project), but are unfamiliar to many social scientists (the closest model would be surveys such as the General Social Survey, the American National Election Studies, and the Panel Study of Income Dynamics). The ultimate success of any such model is therefore subject to its acceptance by the relevant research community, which cannot be guaranteed *ex ante*. Nonetheless, the model has some advantages over the traditional single investigator model that we believe increase its chances of successful adoption.

1) It will enable the research community to better leverage the data assets to produce many times the research output that would be possible with a traditional laboratory model in which both data collection/curation and research are conducted in-house.
2) It lends itself to more comparable research, as researchers can more easily replicate the questions, data, and analytics of previous work, when conducting new inquiries. Often replication efforts are complicated by potentially subtle differences in framing, data, and methods between exploratory and confirmatory studies.
3) It allows researchers to contribute in a variety of ways including a) additional data sources (e.g., text of radio transcripts, social media data); b) improved methods for processing and/or analyzing existing data (e.g., better named entity extraction or topic identification); c) appending useful metadata derived from their own research (e.g., content categories, partisanship labels); d) direct financial support from research grants to support overhead. By accommodating different types of contributions, a shared infrastructure approach should appeal to a wider range of potential collaborators, thereby also increasing its value to subsequent researchers.

Objective 3: Communicate Insights to Nonacademic Stakeholders. An important facilitator of success in the proposed research enterprise is that it be perceived as both legitimate (i.e., rigorous, transparent, and nonpartisan) and also useful. In addition to gathering and organizing data and coordinating research across many research groups, an important goal is therefore to translate the output of the work for nonacademic audiences. More broadly, it is important to advocate for the importance of the social sciences in addressing critical needs, like information ecosystem design in democracies. Although there are many ways to engage stakeholders outside of academia (e.g., blog posts, white papers), one interesting approach that naturally leverages the existence of a centralized data infrastructure is to expose the data itself via web-based interactive visualizations (aka "dashboards") that allow journalists, activists, policymakers, researchers, and members of the public to explore the evidence directly. Another benefit of data dashboards is that, in contrast with published research findings, they are dynamic entities that maintain their relevance even in a fast-moving environment. Rather than reading a statistic about the prevalence of fake news or the diversity of news consumption as it was when the researchers did their work months or even years ago, for example, a dashboard populated with (nearly) live data could show its prevalence as of yesterday, as well as how it has changed in the past week, month, or year. Visualizing data in a way that is psychologically effective and also scientifically valid is a nontrivial undertaking that requires expertise in statistics, user experience design, and software development as well as the substantive domain in question (52, 53). Without downplaying the challenges inherent in designing and implementing useful interactive dashboards, we hope that they will help to ground the public debate around misinformation and democracy on rigorous, nonpartisan evidence.

Objective 4: Develop Academic-Industry Partnerships around Data and Solutions. Modifying the information ecosystem to better support democracy is an example of what has been called solution-oriented social science (40, 54, 55), meaning that it advances fundamental understanding of the social sciences in the course of solving concrete problems of practical interest (56). Rather than pursuing a research agenda based purely on theoretical interest, that is, research should address the concrete challenges confronting the participants (e.g., technology and media companies, fact-checking organizations, scientific societies, etc.) in the information ecosystem. To this end, it is critical to foster academic–industry partnerships with the goal of not only understanding but also improving the information ecosystem.

Partnerships could advance solution-oriented research in a variety of ways, including helping to define the research agenda and specific questions, contributing data, providing analytical tools, translating research findings into design principles, and implementing and testing potential solutions. Journalists and media organizations are perfectly situated to ask questions and provide a platform for disseminating results, while technology firms have data that researchers could use, as well as access to analytical tools. For example, voter files offer ground truth voting behavior (57), search queries correlate with certain offline behaviors (58, 59), and lightweight user actions (e.g., replying, liking, sharing, and commenting) are a useful proxy for engagement. Finally, beyond harvesting existing telemetry data, the capability to design, implement, and test interventions (e.g., reducing uncivil discourse, increasing relative consumption of high-quality information, etc.) requires direct access to proprietary platforms.

The topic of academic–industry partnerships around data has been of increasing interest to academic researchers (see, e.g., ref. 60), but only limited progress has been made in securing the cooperation of industry partners. Perhaps the most prominent recent example is Social Science One (https://socialscience.one/), a commission of senior academics who work with companies (thus far restricted to Facebook) to make preapproved datasets available to researchers while also waiving their right to suppress publication of unfavorable results (39). Although Social Science One is promising, our proposed approach differs from it by starting first with an independent, researcher-designed, and managed data infrastructure. As both these models, along with other models that are being developed in the domain of government administrative data (see, e.g., https://www.aisp.upenn.edu/) and health informatics (see, e.g., https://saildatabank.com/), have their respective strengths and weaknesses, we see them as complements rather than substitutes.

Research Questions

In this section, we briefly summarize a selection of completed, in-progress, or planned research projects that utilize data of the sort described above. These examples are intended only to illustrate

some possibilities and not to limit the scope of the overall research agenda, which we hope will be determined by the collective creativity of a whole research community.

Putting Fake News in Context. As described above, in recent work (18), we have quantified fake news consumption across multiple platforms including television, desktop, and mobile web, finding that it constitutes less than 1/10th of 1% of total daily media consumption, and less than 1% of overall news consumption. Surprisingly, we also find that news consumption in general constitutes a small fraction of overall media consumption (roughly 14%) and is heavily biased toward television across all age categories.

Selection vs. Framing. Which is more important to the underlying and perceived partisanship of publications: selection (which topics they choose to cover) or framing (what slant they give those topics they select to cover)? In future work, we plan to track and map both activities historically and in real time for daily news events spanning television and online content.

Content Overlap in Online News. In response to declining revenue, news publishers have reduced costs by replacing original content with copied or slightly edited versions of generic stories provided via wire services (i.e., AP, Reuters). In ongoing work, we are attempting to quantify the proportion of news reporting that is either copied or unique, as well as the patterns of content overlap that exist within and between news articles. In future work, we will construct networks of publishers characterized by their cocopying patterns, identifying clusters of redundant coverage.

Snippet-Based Content Classification. Prior work on news consumption has relied on classifications of content at the domain (e.g., http://nytimes.com or http://infowars.com) or program (e.g., Today Show, CBS Evening News) level. This approach, while easy to implement, misclassifies content that is not representative of the domain/program of which it is a part (e.g., news content on late-night comedy shows) or is simply not a part of any domain/program (e.g., user-generated content). In ongoing work, we are developing methods using human labelers to classify content at the "snippet" level, where a snippet is defined as a short piece of text or video, thereby allowing us to compare the proportion of news and misinformation across platforms.

Ideologically Segregated Consumption. Partisan echo chambers, and selective exposure to partisan news more generally, are of key concern to communication scholars and the public (61, 62). In ongoing work, we seek to replicate previous findings (63–65) regarding the ideological segregation of online news exposure over the 2016–2018 interval as well as to compare it with television news consumption.

Comparing Survey with Behavioral Data. Surveys are a vital tool in understanding public opinion and knowledge, but have been shown to overestimate news consumption (66, 67). In forthcoming work (68), we show that the bias extends to online and social media-based news consumption and also fails to accurately capture trends. We highlight how behavioral data are more easily adaptable to the wide range of possible results that a researcher may need to answer with different, but related, sets of questions about news consumption.

Measuring Awareness and Understanding of News Events. In ongoing work, we are pulling the top facts from online articles each day and running regular polls that ask 1) whether respondents are aware of a given event, and 2) if so, whether or not they know the facts in question. In addition to measuring the relationship between news coverage and public awareness, this dataset will initiate a larger program of tracking which types of information are absorbed by the news consuming public, and via which channels.

Conclusion

The debate around misinformation and its potentially damaging effects on public opinion, understanding, and democratic decision making is complex and multifaceted. There is not, to our knowledge, any general consensus on what "the problem" is, and even less agreement on what the solution or solutions ought to be (2, 4, 5, 13, 16, 17, 24, 29). We do not pretend that our approach will resolve these disagreements over what matters and what to do about it. To the extent that such disagreements arise and persist because of the absence of systematic empirical evidence, however, we hope that it will help, in two ways. First, the creation of a shared, open data infrastructure to support research on misinformation and its effect on democracy will reduce existing barriers to producing rigorous, replicable, and ultimately useful science. Second, exposing the data and research insights to external stakeholders via continuously updating interactive visualizations will force interlocutors to confront the world as it is (or at least as it has been measured) rather than how they imagine it to be. Of course, we acknowledge that measurement itself is also imperfect in important ways; however, we do not see these shortcomings as a reason not to rely on data, but rather as a motivation to design better instruments and to collect better data. That data will also be imperfect, and the process of discovering that will in turn motivate better instruments, and so on. Just as no one experiment can settle any complex social scientific question, no one dataset can ever satisfactorily capture everything that we might care about. The process of informing our understanding of the world with evidence will therefore be an ongoing one. Our proposal is simply that we cannot afford not to begin this process.

Data Availability. There are no data underlying this work.

ACKNOWLEDGMENTS. We are grateful to Harmony Labs for engineering support, to The Nielsen Company for providing data, and to the Nathan Cummings Foundation and the Carnegie Corporation for seed funding. Finally, Jennifer Allen, Lilia Chang, Anna Croley, Ling Dong, Baird Howland, Homa Hosseinmardi, Rachel Leong, Daniel Muise, and Marcel Wittich have all made valuable contributions to the research.

1. H. Allcott, M. Gentzkow, Social media and fake news in the 2016 election. *J. Econ. Perspect.* **31**, 211–236 (2017).
2. C. Wardle, H. Derakhshan, "Information disorder: Toward an interdisciplinary framework for research and policymaking" (DGI(2017)09, Council of Europe, 2017).
3. K. Shu, A. Sliva, S. Wang, J. Tang, H. Liu, Fake news detection on social media: A data mining perspective. *SIGKDD Explor.* **19**, 22–36 (2017).
4. J. Tucker *et al.*, Social media, political polarization, and political disinformation: A review of the scientific literature. *SSRN* [Preprint] (2018). https://doi.org/10.2139/ssrn.3144139 (Accessed 12 March 2021).
5. D. M. J. Lazer *et al.*, The science of fake news. *Science* **359**, 1094–1096 (2018).
6. S. Vosoughi, D. Roy, S. Aral, The spread of true and false news online. *Science* **359**, 1146–1151 (2018).
7. C. Shao *et al.*, Anatomy of an online misinformation network. *PLoS One* **13**, e0196087 (2018).
8. A. Guess, B. Nyhan, J. Reifler, *Selective Exposure to Misinformation: Evidence from the Consumption of Fake News during the 2016 US Presidential Campaign* (European Research Council, 2018).
9. M. Stella, E. Ferrara, M. De Domenico, Bots increase exposure to negative and inflammatory content in online social systems. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 12435–12440 (2018).
10. N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, D. Lazer, Fake news on Twitter during the 2016 U.S. presidential election. *Science* **363**, 374–378 (2019).
11. A. Guess, J. Nagler, J. Tucker, Less than you think: Prevalence and predictors of fake news dissemination on Facebook. *Sci. Adv.* **5**, eaau4586 (2019).
12. G. Pennycook, D. G. Rand, Fighting misinformation on social media using crowdsourced judgments of news source quality. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 2521–2526 (2019).

13. S. Aral, D. Eckles, Protecting elections from social media manipulation. *Science* **365**, 858–861 (2019).
14. C. Silverman, This analysis shows how viral fake election news stories outperformed real news on Facebook. *BuzzFeed News*, 16 November 2016. https://www.buzzfeednews.com/article/craigsilverman/viral-fake-election-news-outperformed-real-news-on-facebook. Accessed 12 March 2021.
15. J. Roozenbeek, S. van der Linden, The fake news game: Actively inoculating against the risk of misinformation. *J. Risk Res.* **22**, 570–580 (2019).
16. D. J. Watts, D. Rothschild, Don't blame the election on fake news. Blame it on the media. *Columbia Journalism Review*, 5 December 2017. https://www.cjr.org/analysis/fake-news-media-election-trump.php. Accessed 12 March 2021.
17. B. Nyhan, Why fears of fake news are overhyped. *Medium*, 4 February 2019. https://link.medium.com/MgiT8vLJUW. Accessed 12 March 2021.
18. J. Allen, B. Howland, M. Mobius, D. Rothschild, D. J. Watts, Evaluating the fake news problem at the scale of the information ecosystem. *Sci. Adv.* **6**, eaay3539 (2020).
19. B. Bago, D. G. Rand, G. Pennycook, Fake news, fast and slow: Deliberation reduces belief in false (but not true) news headlines. *J. Exp. Psychol. Gen.* **149**, 1608–1613 (2020).
20. W. P. Eveland, M. Seo, K. Marton, Learning from the news in campaign 2000: An experimental comparison of TV news, newspapers, and online news. *Media Psychol.* **4**, 353–378 (2002).
21. J. H. Walma van der Molen, T. H. A. van der Voort, Children's recall of television and print news: A media comparison study. *J. Educ. Psychol.* **89**, 82–91 (1997).
22. M. Dijkstra, H. E. J. J. M. Buijtels, W. F. van Raaij, Separate and joint effects of medium type on consumer responses: A comparison of television, print, and the Internet. *J. Bus. Res.* **58**, 377–386 (2005).
23. M. Dijkstra, W. F. van Raaij, Media effects by involvement under voluntary exposure: A comparison of television, print and static internet. *J. Euro-Mark* **11**, 1–21 (2002).
24. T. E. Patterson, *How America Lost its Mind: The Assault on Reason That's Crippling Our Democracy* (University of Oklahoma Press, 2019).
25. D. Miller, *Tell Me Lies: Propaganda and Media Distortion in the Attack on Iraq*, D. Miller, Ed. (Pluto Press, 2003).
26. W. Lance Bennett, R. G. Lawrence, S. Livingston, *When the Press Fails: Political Power and the News Media from Iraq to Katrina* (University of Chicago Press, 2008).
27. B. Nyhan, Why the "death panel" myth wouldn't die: Misinformation in the health care reform debate. *The Forum*, 10.2202/1540-8884.1354 (2011).
28. R. G. Lawrence, M. L. Schafer, Debunking Sarah Palin: Mainstream news coverage of "death panels.". *Journalism* **13**, 766–782 (2012).
29. Y. Benkler, R. Faris, H. Roberts, *Network Propaganda: Manipulation, Disinformation, and Radicalization in American Politics* (Oxford University Press, 2018).
30. T. Rogers, R. Zeckhauser, F. Gino, M. I. Norton, M. E. Schweitzer, Artful paltering: The risks and rewards of using truthful statements to mislead others. *J. Pers. Soc. Psychol.* **112**, 456–473 (2017).
31. AllSides, How to spot 11 types of media bias. https://www.allsides.com/media-bias/how-to-spot-types-of-media-bias. Accessed 12 March 2021.
32. S. Mullainathan, A. Shleifer, The market for news. *Am. Econ. Rev.* **95**, 1031–1053 (2005).
33. M. Gentzkow, J. M. Shapiro, What drives media slant? Evidence from US daily newspapers. *Econometrica* **78**, 35–71 (2010).
34. B. G. Southwell, E. A. Thorson, L. Sheble, *Misinformation and Mass Audiences* (University of Texas Press, 2018).
35. S. Dentzer, Communicating medical news—pitfalls of health care journalism. *N. Engl. J. Med.* **360**, 1–3 (2009).
36. W. Glazer, Scientific journalism: The dangers of misinformation. *Curr. Psychiatr.* **12**, 33–35 (2013).
37. D. A. Scheufele, N. M. Krause, Science audiences, misinformation, and fake news. *Proc. Natl. Acad. Sci. U.S.A.* **116**, 7662–7669 (2019).
38. P. Rosenzweig, *The Halo Effect* (Free Press, 2007).
39. G. King, N. Persily, A new model for industry-academic partnerships. *PS (Wash DC)* **53**, 703–709 (2018).
40. D. J. Watts, Should social science be more solution-oriented? *Nat. Hum. Behav.* **1**, 0015 (2017).
41. W. Wang, R. Kennedy, D. Lazer, N. Ramakrishnan, Growing pains for global monitoring of societal events. *Science* **353**, 1502–1503 (2016).
42. D. C. Mutz, *Hearing the Other Side: Deliberative Versus Participatory Democracy* (Cambridge University Press, 2006).
43. D. M. Kahan *et al*., The polarizing impact of science literacy and numeracy on perceived climate change risks. *Nat. Clim. Chang.* **2**, 732–735 (2012).
44. C. A. Bail *et al*., Exposure to opposing views on social media can increase political polarization. *Proc. Natl. Acad. Sci. U.S.A.* **115**, 9216–9221 (2018).
45. W. Wang, D. Rothschild, S. Goel, A. Gelman, Forecasting elections with non-representative polls. *Int. J. Forecast.* **31**, 980–991 (2015).
46. A. Gelman, S. Goel, D. Rivers, D. Rothschild, The mythical swing voter. *Quart. J. Polit. Sci.* **11**, 103–130 (2016).
47. D. J. Watts, Computational social science: Exciting progress and future directions. *Bridge Front. Eng.* **43**, 5–10 (2013).
48. C. Wells, K. Thorson, Combining big data and survey techniques to model effects of political content flows in Facebook. *Soc. Sci. Comput. Rev.* **35**, 33–52 (2017).
49. T. Konitzer, D. Rothschild, S. Hill, K. C. Wilbur, Using big data and algorithms to determine the effect of geographically targeted advertising on vote intention: Evidence from the 2012 US presidential election. *Polit. Commun.* **36**, 1–16 (2019).
50. C. Budak, D. J. Watts, Dissecting the spirit of Gezi: Influence vs. selection in the occupy Gezi movement. *Sociol. Sci.* **2**, 370–397 (2015).
51. M. J. Salganik, I. Lundberg, A. T. Kindel, S. McLanahan, Introduction to the special collection on the fragile families challenge. *Socius*, 10.1177/2378023119871580 (2019).
52. J. Mackinlay, Automating the design of graphical presentations of relational information. *ACM Trans. Graph.* **5**, 110–141 (1986).
53. M. Hegarty, The cognitive science of visual-spatial displays: Implications for design. *Top. Cogn. Sci.* **3**, 446–474 (2011).
54. M. Prasad, Problem-solving sociology. *Trajectories* **28**, 17–21 (2016).
55. M. Western, "We need more solution-oriented social science: On changing ourframes of reference and tackling big social problems". *Impact of Social Sciences Blog*, 26 June 2016. http://eprints.lse.ac.uk/67288/. Accessed 12 March 2021.
56. D. E. Stokes, *Pasteur's Quadrant: Basic Science and Technological Innovation* (Brookings Institution Press, Washington, DC, 1997).
57. B. C. Burden, Voter turnout and the national election studies. *Polit. Anal.* **8**, 389–398 (2000).
58. S. Goel, J. M. Hofman, S. Lahaie, D. M. Pennock, D. J. Watts, Predicting consumer behavior with Web search. *Proc. Natl. Acad. Sci. U.S.A.* **107**, 17486–17490 (2010).
59. H. Choi, H. Varian, Predicting the present with Google trends. *Econ. Rec.* **88**, 2–9 (2012).
60. SSRC, To secure knowledge: Social science partnerships for the common good. https://www.ssrc.org/to-secure-knowledge/. Accessed 12 March 2021.
61. E. Pariser, *The Filter Bubble: What the Internet Is Hiding from You* (Penguin, 2011).
62. C. R. Sunstein, *Republic: Divided Democracy in the Age of Social Media* (Princeton University Press, 2018).
63. M. Gentzkow, J. M. Shapiro, "Ideology and online news" in *Economic Analysis of the Digital Economy*, A. Goldfarb, S. M. Greenstein, C. E. Tucker, Eds. (University of Chicago Press, 2015), pp. 169–190.
64. E. Bakshy, S. Messing, L. A. Adamic, Political science. Exposure to ideologically diverse news and opinion on Facebook. *Science* **348**, 1130–1132 (2015).
65. S. Flaxman, S. Goel, J. M. Rao, Filter bubbles, echo chambers, and online news consumption. *Public Opin. Q.* **80**, 298–320 (2016).
66. M. Prior, The immensely inflated news audience: Assessing bias in self-reported news exposure. *Public Opin. Q.* **73**, 130–143 (2009).
67. M. Prior, The challenge of measuring media exposure: Reply to Dilliplane, Goldman, and Mutz. *Polit. Commun.* **30**, 620–634 (2013).
68. T. Konitzer *et al*., Comparing estimates of news consumption from survey and passively collected behavioral data. *SSRN* [Preprint] (2020). https://doi.org/10.2139/ssrn.3548690 (Accessed 12 March 2021).