To what extent are democratic elections vulnerable to social media manipulation? The fractured state of research and evidence on this most important question facing democracy is reflected in the range of disagreement among experts. Facebook chief executive officer Mark Zuckerberg has repeatedly called on the U.S. government to regulate election manipulation through social media. But we cannot manage what we do not measure. Without an organized research agenda that informs policy, democracies will remain vulnerable to foreign and domestic attacks. Thankfully, social media’s effects are, in our view, eminently measurable. Here, we advocate a research agenda for measuring social media manipulation of elections, highlight underutilized approaches to rigorous causal inference, and discuss political, legal, and ethical implications of undertaking such analysis. Consideration of this research agenda illuminates the need to overcome important trade-offs for public and corporate policy—for example, between election integrity and privacy.

We have promising research tools, but they have not been applied to election manipulation, mainly because of a lack of access to data and lack of cooperation from the platforms (driven in part by public policy and political constraints).

Two recent studies commissioned by the U.S. Senate Intelligence Committee detail Russian misinformation campaigns targeting hundreds of millions of U.S. citizens during the 2016 presidential election. The reports highlight, but do not answer,
For example, unlike the majority of linear television advertising, social media can be personally targeted; assessing its reach requires analysis of paid and unpaid media, ranking algorithms and advertising auctions; and causal analysis is necessary to understand how social media changes opinions and voting.

Luckily, much of the necessary methodology has already been developed. A growing body of literature illuminates how social media influences behavior. Analysis of misinformation on Twitter and Facebook (4, 5), and randomized and natural experiments involving hundreds of millions of people on various platforms, have shown how social media changes what we shop, read, and exercise [e.g., (6, 7)]. Similar methods can and should be applied to voting (8).

Research on election manipulation will be enabled and constrained by parallel policy initiatives that aim, for example, to protect privacy. Although privacy legislation may prohibit retention of consumer data, such data may also be critical to understanding how to harden our democracies against manipulation. To preserve democracy in the digital age, we must manage these trade-offs and overcome multidisciplinary methodological challenges simultaneously.

MEASURING MANIPULATION

We propose a four-step research agenda for estimating the causal effects of social media manipulation on voter turnout and vote choice (see the figure). We also describe analysis of the indirect, systemic effects of social media manipulation on campaign messaging and the news cycle (see supplementary materials for further details).

Step 1: We must catalog exposures to manipulation, which we define as impressions (i.e., serving of an ad or message to a viewer) of paid and organic manipulative content (9) (e.g., false content intended to deceive voters, or even true content propagated by foreign actors, who are banned from participating in domestic political processes, with the intent of manipulating voters). To do so, we must evaluate the reach of manipulation campaigns and analyze the targeting strategies that distribute these impressions. For example, we need to know which text, image, and video messages were advertised, organically posted, and “boosted” through paid advertising, and on which platforms, as well as when and how each of these messages was shared and reshared by voters (2) and authentic accounts. Here, understanding social multiplier effects, or how individuals influence each other, will be essential, and the literature on peer effects in social networks describes how our peers change our behavior (6–8). The content of the messages should also be analyzed to assess the effectiveness of particular textual, image, and video content in changing opinions and behavior.

Much prior work on exposure to and diffusion of (mis)information has relied on proxies for exposure, such as who follows whom on social media (2, 4), though some has also investigated logs of impressions, recognizing the role of algorithmic ranking and auctions in determining exposure [e.g., (5, 10)]. Given prior work on the rapid decay of advertising effects, it is important to consider when these exposures occurred, as recent work suggests that exposure to misinformation may increase just prior to an election and wane immediately afterward (2).

Step 2: We must combine exposure data with data on voting behavior. Data about voter turnouts in the United States are readily available in public records (e.g., registered voters’ names, addresses, party affiliations, and when they voted). Prior work has matched social media accounts and public voting records using relatively coarse data (e.g., residences inferred from self-reported profile data and group-level, anonymous matching procedures) (2, 8), in part because of privacy concerns, resulting in low match rates that limit statistical power and representativeness. This could be substantially improved, for example, by using the rich location data possessed by social media platforms, similar to that already sold and reused for marketing purposes (e.g., matching voter registrations with inferred home addresses based on mobile and other location data), rather than simply matching voters by name and age at the state level.

In contrast to turnout data, vote choices in the United States are secret and thus only measurable in aggregate (e.g., precinct-level vote totals and shares) or sparsely and indirectly through surveys (e.g., exit polls). Thus, exposure data would need to be aggregated, at the precinct, district, or state levels, before combining it with vote choice data, making it likely that estimates of voter turnout effects will be more precise than estimates of vote choice effects.

Experiments demonstrate that persuasive interventions can substantially affect voter turnout. But, when assessing turnout, it is important to remember that voting is habitual. Effective manipulation therefore likely requires targeting occasional voters in battleground regions. In social media, however, this type of targeting is possible and took place during the 2016 U.S. presidential election. Analysis of the precision of targeting efforts is essential to understanding voter turnout effects.

Influencing vote choice is more difficult because likely voters have strong prior beliefs. However, even the pessimistic litera-
A blueprint for empirical investigations of social media manipulation

<table>
<thead>
<tr>
<th>ASSESS MESSAGE CONTENT AND REACH</th>
<th>ASSESS TARGETING AND EXPOSURE</th>
<th>ASSESS CAUSAL BEHAVIOR CHANGE</th>
<th>ASSESS EFFECTS ON VOTING BEHAVIOR</th>
</tr>
</thead>
<tbody>
<tr>
<td>How many messages spread?</td>
<td>Who was exposed to which messages?</td>
<td>How did messages change opinions and behavior?</td>
<td>How did opinion and behavior change after exposure?</td>
</tr>
<tr>
<td>Analysis of paid and organic information diffusion</td>
<td>Analysis of targeting and messaging exposure</td>
<td>Causal statistical analysis of opinion and behavior change</td>
<td>Counterfactual analysis of deviations from expected voting</td>
</tr>
<tr>
<td>Measure impressions through paid media and sharing</td>
<td>Evaluate targeting campaigns and impression distributions</td>
<td>Evaluate causal effects across individuals and segments</td>
<td>Measure deviations from expected voting behavior</td>
</tr>
</tbody>
</table>

A routine test of the kind these platforms conduct every day, can be used to estimate the effects of exposure to such content (13). Such experiments could facilitate measurement of both direct effects (e.g., effects of manipulative content on recipients) and indirect “spillover” effects (e.g., word of mouth from recipients to peers), though other methods for estimating the latter also exist (6–8).

One important challenge is that statistical precision is often inadequate to answer many questions about effects on voter behavior. For example, randomized experiments conducted by Facebook in the 2010 and 2012 U.S. elections only barely detected effects on turnout—even though the estimated effects imply that a minimal intervention caused hundreds of thousands of additional votes to be cast [e.g., (8)]. The lack of statistical precision in those studies arose in part because only about a tenth of users were uniquely matched to voter records, which, as we note, could be improved upon. Furthermore, unlike television advertising, much less of the as-good-as-random variation in exposure to social media may be within, not between, geographic areas, making effects on aggregate vote shares more difficult to detect. Such imprecision can be misleading, suggesting that online advertising does not work simply because the effects were too small to detect in a given study (14), even though the results were consistent with markedly low costs per incremental vote, making engagement in such campaigns economically rational.

Step 4: We must compute the aggregate consequences of changes in voting behavior for election outcomes. To do so, we would combine summaries of individual-level counterfactuals (i.e., predicted voter behavior with and without exposure) with data on the abundance of exposed voters by geographic, demographic, and other characteristics in specific elections. This would enable estimates and confidence intervals for vote totals in specific states or regions if a social media manipulation campaign had not been conducted. Although some of these confidence intervals will include vote totals that do or do not alter the winner in a particular contest, the ranges of counterfactual outcomes would still be informative about how such manipulation can alter elections. Although it remains to be seen exactly how precise the resulting estimates of the effects of exposure to misinformation would be, even sufficiently precise and carefully communicated null results could exclude scenarios currently posited by many commentators.

Research should also address the systemic effects of social media manipulation, like countermessaging and feedback on the news cycle itself. Countermessaging could be studied in, for example, the replies to and debunking of fake news on Facebook and Twitter (4, 5) and whether the emergence of fake stories alters the narrative trajectories of messaging by campaigns or other interested groups. Feedback into the news cycle could be studied by examining the causal impact of manipulation on the topical content of news coverage. For example, Ananya Sen and Pinar Yildirim have used as-good-as-random variation in the weather to show that more viewership to particular news stories causes publishers to write more stories on those topics. A similar approach could determine whether attention to misinformation alters the topical trajectory of the news cycle.

We believe near-real-time and ex post analysis are both possible and helpful. The bulk of what we are proposing is ex post analysis of what happened, which can then be used to design platforms and policy to...
We have described what a rigorous analysis of social media manipulation would entail, but have also assumed that the data required to conduct it are available for analysis. But does the social media data that we describe above, especially data about the content that individuals were exposed to, exist retrospectively or going forward? Social media companies routinely log what users are exposed to for research and retraining algorithms. But current regulatory regimes disincentivize the lossless retention of this data. For example, the European Union’s General Data Protection Regulation (GDPR) encourages firms to comply with user requests to delete data about them, including content that they have posted. An audit by the office of the Irish Data Protection Commissioner caused Facebook to implement similar policies in 2012. Thus, without targeted retention, it may be difficult for firms to accurately quantify exposures for users who deleted their accounts or were exposed to content deleted by others. We should recognize that well-intentioned privacy regulations, though important, may also impede assessments like the one that we propose. Similarly, proposed legislation in the United States (the DETOUR Act) could make many routine randomized experiments by these firms illegal (Senate Bill 1084), making future retrospective analyses more difficult and, of course, making ongoing efforts by those firms to limit such manipulation less data-driven.

Even if such data are available, it is not obvious that we should accept world governments demanding access to or analyses of those data to quantify the effects of speech in elections. Although we suggest that linking datasets could be achieved using rich location data routinely used for marketing, such use may be reasonably regarded as data misuse. Thus, we do not unconditionally advocate the use of any and all existing data for the proposed analyses. Instead, privacy-preserving methods for record linkage and content analysis, such as differential privacy (15), could help manage trade-offs between the need for privacy and the need to protect democracy.

Hardening democracies to manipulation will take extraordinary political and commercial will. Politicians in the United States, for example, may have counter-vailing incentives to support or oppose a postmortem on Russian interference, and companies like Facebook, Twitter, and Google face pressure to secure personal data. Perhaps this is why Social Science One, the forward-looking industry–academic partnership working to provide access to funding and Facebook data to study the effects of social media on democracy, faced long delays in securing access to any data, and why its most recent release does not include any data relevant to a postmortem on Russian interference in the 2016 or 2018 elections in the United States. Moreover, this cannot just be about any single company or platform. Comprehensive analysis must include Facebook, Twitter, YouTube, and others. Perhaps only mounting pressure from legislators and the public will empower experts with the access they need to do the work that is required.

Research collaborations with social media platforms, like that being undertaken by Social Science One, can facilitate access to important data for understanding democracy’s vulnerability to social media manipulation. We hope the realization that the analysis we propose is bigger than any one election and essential to protecting democracies worldwide will help overcome partisanship and myopic commercial interests in making the necessary data available, in privacy-preserving ways.

However, it is important to note that prior work has linked social media messaging to validated voting, both with the assistance of the social media platforms (8) and without it (2). Although collaboration with the platforms is preferable, it is not the only way to assess manipulation. In the absence of commercial or governmental support for postmortems on past elections, active analysis of ongoing information operations, conducted according to the framework that we propose, is a viable and valuable alternative. A detailed understanding of country-specific regulations and election procedures is necessary for robust analysis of the effects of social media manipulation on democracies worldwide.

Our suggested approach emphasizes precise causal inference, but this should be complemented with surveys, ethnographies, and analysis of observational data to understand the mechanisms through which manipulation can affect opinions and behavior.

Achieving a scientific understanding of the effects of social media manipulation on elections is an important civic duty. Without it, democracies remain vulnerable. The sooner we begin a public discussion of the trade-offs between privacy, free speech, and democracy that arise from the pursuit of this science, the sooner we can realize a path forward.

ACKNOWLEDGMENTS
We thank A. J. Berinsky and B. Nyhan for comments. S.A. has financial interest in Alibaba, Google, Amazon, and Twitter. S.A. was a Scholar in Residence at the University of California, Irvine, in 2016–2017. D.E. has financial interest in Microsoft, Google, Amazon, and Twitter. S.A. has received research funding from Harvard University, Google, and the National Science Foundation. D.E. is a visiting researcher at Microsoft in 2016 and has received funding from Amazon, Google, and the National Science Foundation.

SUPPLEMENTARY MATERIALS
science.sciencemag.org/content/365/6456/858/suppl/DC1
10.1126/science.aaw8243
Protecting elections from social media manipulation
Sinan Aral and Dean Eckles

Science 365 (6456), 858-861.
DOI: 10.1126/science.aaw8243
Supplementary Materials for

Protecting elections from social media manipulation

Sinan Aral* and Dean Eckles

*Corresponding author. Email: sinan@mit.edu

Published 30 August 2019, Science 365, 858 (2019)
DOI: 10.1126/science.aaw8243

This PDF file includes:

Supplementary Text
References
Supplementary Text

Here we provide additional references supporting key arguments in the main text.

1. **DISAGREEMENT AMONGST EXPERTS**


Disagreement among experts about whether social media manipulation has or could affect the results of elections stems from differing beliefs about (a) the likely reach and scope of misinformation campaigns and (b) the likely effects of social media manipulation on voter turnout and vote choice.

2. **THE REACH AND SCOPE OF MISINFORMATION CAMPAIGNS**

While some research estimates that Russian misinformation, for example, reached hundreds of millions of people on social media during the 2016 U.S. Presidential election (DiResta et al., 2018; Howard et al., 2018), others contend the reach and scope of exposures was small, concentrated and selective (Allcott & Gentzkow, 2017; Grinberg et al., 2019, Guess, Nagler & Tucker, 2019; Guess, Nyhan & Reifler, 2018).

3. **EFFECTS ON VOTER TURNOUT AND VOTE CHOICE**

There is also disagreement on the effectiveness of social media persuasion and whether it could be substantial enough to tip an election. Here it is important to distinguish the likely effects of manipulation on voter turnout and vote choice.

With regard to vote choice, some meta-analytic reviews suggest the effects of impersonal contact (e.g., mailing, TV and digital advertising) on vote choice in elections are very small. For example, Kalla and Broockman (2017) conclude that “the best estimate of the size of persuasive effects [i.e., effects of advertising on vote choice] in general elections in light of our evidence is zero.” However, there remains substantial uncertainty and heterogeneity in their estimates, reflected in the confidence interval reported in the main text, from their Figure 4b, for the meta-analytic effect of impersonal contact within two months of election day. Kalla and Broockman (2017) also find significant meta-analytic effects on vote choice in primaries, issue specific ballot measures and when campaigns target persuadable voters, suggesting the possibility that manipulation is effective in changing vote choices when they are issue specific and targeted. We also note that the social media manipulation we have observed to date has typically been issue specific and targeted, similarly to the randomized intervention cited in the main text (Rogers & Nickerson, 2013).
Furthermore, social media manipulation does not have to affect vote choice to tip an election. Effects on voter turnout, if well targeted, could be substantial enough to change an overall result. The meta-analytic assessments of voter turnout point to much more substantial effects. For example, the meta-analysis by Green et al. (2013) estimates that direct mailings with social pressure generate an average increase in voter turnout of 2.9% (95% CI = 2.7%-3.0%), canvassing generates an average increase of 2.5% (95% CI = 1.8%-3.3%) and volunteer phone banks generate an average increase of 2% (95% CI = 1.3%-2.6%). Dale and Strauss (2009) estimate the voter turnout effect of text messages to be 4.1% and there is also evidence that personalized emails create substantial voter turnout effects (Davenport 2012; Malhotra et al., 2012). The only studies of voter turnout effects from social media messaging estimate that hundreds of thousands of additional votes were cast as a result of social media messages (Bond et al., 2012; Jones et al., 2017).

4. MEASURING EXPOSURE

Much prior work on exposure to and diffusion of (mis)information has relied on proxies for exposure. However, some work by researchers at Facebook (Bakshy et al., 2012a,b; Bakshy, Eckles & Bakshy, 2017; Friggeri et al., 2014; Messing & Adamic, 2015; Messing, 2013) has made use of detailed data about impressions (delivery of content to the users’ device), including information about what content was actually displayed to a user for at least a minimum period of time, thus making use of measures now in widespread use in digital advertising.

5. LINKING EXPOSURE TO VOTING DATA

Voter turnout in the United States is a matter of public record, so voter records including data about individuals’ turnout is widely used by campaigns and researchers. Bond et al. (2012) linked Facebook accounts to voter turnout data to estimate effects of a randomized intervention. They did this matching using limited information, apparently because of privacy concerns, as articulated in a companion paper about privacy-preserving record linkage (Jones et al., 2013). A subsequent experiment used similarly coarse data for record linkage and resulted in similarly low unique match rates (Jones et al., 2017).

6. BIAS IN NAÏVE OBSERVATIONAL STUDIES

Aral, Muchnik, and Sundararajan (2009) compare the results of naïve observational methods to counterfactual methods based on matching and find naïve methods overestimate the effects of non-paid exposure to behavior of friends in an online social network by 300-700%. At least since LaLonde (1986), researchers have used randomized experiments as a “gold standard” with which to evaluate other, observational methods, like matching. In the context of the diffusion of (mis)information, Eckles and Bakshy (2017) validate the methods used in Aral, Muchnik and Sundararajan (2009) by comparing the results of a large field experiment on Facebook to analyses of matching methods. They find naïve methods overestimate the effects of non-paid exposure to content shared by friends by over 300% and demonstrate that matching can reduce this bias by up to 80–100%.
Non-experimental methods for estimating effects of paid exposure (digital advertising) have also performed poorly when similarly evaluated. Gordon et al. (2018) used randomized experiments to show observational estimates of social media influence, without careful causal inference, are frequently off by over 100%.

Similar confounding is plausibly present in widely-publicized claims (Matz et al., 2018) about the effectiveness of targeting ads according to inferred personality traits (Eckles, Gordon & Johnson, 2018).

7. QUASI-EXPERIMENTAL METHODS FOR ESTIMATING EFFECTS ON VOTING BEHAVIORS

Several studies have exploited a mismatch between borders of competitive electoral districts and borders of regions for marketing purposes to study effects of advertising, including Huber and Arceneaux (2007) and more recent work (Spenkuch & Toniatti, 2018; Wang, Lewis & Schweidel, 2018). However, targeting of digital advertising is less restricted to such borders compared with traditional, linear television advertising, making this source of plausibly exogenous variation in exposure largely inapplicable in the digital arena.

8. ROUTINE EXPERIMENTATION BY PLATFORMS

Internet companies are engaged in continual experimentation, with the most prominent firms starting hundreds of experiments each week (Bakshy, Eckles & Bernstein, 2014; McAfee, A., & Brynjolfsson, 2012; Varian, 2016). Even a single part of a product, such as the algorithm for ranking search results or a feed of content shared by others (e.g., News Feed), might be modulated in hundreds or thousands of experiments over the course of a campaign (Kohavi & Thomke, 2017; Peysakhovich & Eckles, 2018). Most of these experiments are not designed for studying exposure to political content, with the exception of, e.g., Messing (2013), covered in Sifry (2014). However, such experiments can be key inputs for recently developed methods for high-dimensional instrumental variables regression (e.g., Kang et al., 2016; Belloni et al., 2017; Peysakhovich & Eckles, 2018; Guo et al., 2018).

9. INDIRECT EFFECTS

Like other content on social media, effects of Russian-sponsored content may occur indirectly via diffusion of the content and further social contagion (cf. Nickerson, 2008; Bond et al., 2012; Jones et al., 2017). There has been substantial recent development of methods for estimation of such "spillover" effects in networks (e.g. Aronow, 2012; Aronow & Samii, 2017; Athey, Eckles & Imbens, 2018), with empirical work in online social networks making use of both designed experiments (e.g., Aral & Walker, 2011, 2012, 2014; Bakshy et al., 2012a,b; Eckles, Kizilec & Bakshy, 2016; Huang et al., 2019; Muchnik, Aral & Taylor, 2013), natural quasi-experiments (e.g., Aral & Nicolaides, 2017; Aral & Zhao, 2018) and other causal inference methods (e.g., Aral, Muchnik & Sundararajan, 2009; Eckles & Bakshy, 2017).
There is also reason to believe that there are temporal spillovers, with the effects of persuasive messaging in one election spilling over into future elections (Gerber, Green & Shachar, 2003; Davenport et al., 2010; Bedolla & Michelson, 2012). This is consistent with the idea that voting is habitual (e.g., Plutzer, 2002; Gerber, Green & Shachar, 2003) and that messaging can affect voting habits.

To the extent that social media activity is subsequently covered by the news media, this might result in effects on the voting behavior of those who are not directly exposed on social media. This would require different empirical strategies for credible causal inference, such as in Sen and Yildirim (2016).

References


Messing, S., & Westwood, S. J. (2014). Selective exposure in the age of social media: Endorsements trump partisan source affiliation when selecting news online. Communication Research, 41(8), 1


Senate Bill 1084 (2019). Deceptive Experiences To Online Users Reduction Act, 116th Congress.