INSIGHTS

POLICY FORUM

SCIENCE AND DEMOCRACY

Protecting elections from social media manipulation

Rigorous causal analysis could help harden democracy against future attacks

By **Sinan Aral**[1,2,3] *and* **Dean Eckles**[1,2]

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,

[1]*Sloan School of Management, Massachusetts Institute of Technology (MIT), Cambridge, MA, USA.* [2]*Institute for Data, Systems, and Society, MIT, Cambridge, MA, USA.* [3]*Manifest Investment Partners, Tiburon, CA, USA. Email: sinan@mit.edu*

ILLUSTRATION: BROBEL DESIGN

whether social media manipulation may have influenced the outcome. Some experts argue that Russia-sponsored content on social media likely did not decide the election because Russian-linked spending and exposure to fake news (*1*, *2*) were small-scale. Others contend that a combination of Russian trolls and hacking likely tipped the election for Donald Trump (*3*). Similar disagreements exist about the UK referendum on leaving the European Union and recent elections in Brazil, Sweden, and India.

Such disagreement is understandable, given the distinctive challenges of studying social media manipulation of elections. 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 how 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 tradeoffs 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 inauthentic 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 turnout 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-

ture on vote choice allows for substantial effects, especially when targeted messages change voters' beliefs. In a meta-analysis of 16 field experiments, Kalla and Broockman (*11*) report a wide 95% confidence interval (CI) of [−0.27%, 0.83%] for the effect of impersonal contact (e.g., mail, ads) on vote choice within 2 months of general elections, and larger, more significant effects in primaries and on issue-specific ballot measures. In Rogers and Nickerson (*12*), informing recipients in favor of abortion rights that a candidate was not consistently supportive of such rights had a 3.90% [95% CI: 1.16%, 6.64%] effect on reported vote choice. Such prior beliefs are predictable and addressable in manipulation campaigns through social media targeting and thus measurable in studies of the effectiveness of such manipulation.

Step 3: We must assess the effects of manipulative messages on opinions and behavior. This requires a rigorous approach to causal inference, as naïve, observational approaches would neglect the confounding factors that cause both exposure and voting behavior (e.g., voters targeted with such content are more likely to be sympathetic to it). Evaluations using randomized experiments have shown that observational estimates of social media influence without careful causal inference are frequently off by more than 100%. Effects of nonpaid exposures, estimated without causal inference, have been off by as much as 300 to 700%. Yet, causal claims about why social media messages spread are routinely made without any discussion of causal inference. Widely publicized claims about the effectiveness of targeting voters by inferred personality traits, as allegedly conducted by Cambridge Analytica, were not based on randomized experiments or any other rigorous causal inference and therefore plausibly suffer from similar biases.

To credibly estimate the effects of misinformation on changes in opinions and behaviors, we must change our approach and embrace causal inference. We must analyze similar people exposed to varying levels of misinformation, perhaps due to random chance or explicit randomization by firms and campaigns. Fortunately, there are many, until-now largely ignored, sources of such random variation. For example, Facebook and Twitter constantly test new variations on their feed ranking algorithms, which cause people to be exposed to varying levels of different types of content. Some preliminary analysis suggests that an A/B test run by Facebook during the 2012 U.S. presidential election caused over 1 million Americans to be exposed to more "hard news" from established sources, affecting political knowledge, policy preferences, and voter turnout (*10*). Most of these routine experiments are not intended specifically to modulate exposure to political content, but recent work has illustrated how the random variation produced by hundreds or thousands of routine tests, 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

A blueprint for empirical investigations of social media manipulation

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

prevent future manipulation. The pace at which voting data (whether in primaries or general elections) become available is a key limitation. But real-time detection of manipulation efforts and reaction to them could also be designed, similar to tactics in digital advertising that estimate targeting models offline and then implement real-time bidding based on those estimates. Experimental analysis of the effect of social media on behavior change can be spun up and conducted by the platforms in a matter of days and analyzed in a week.

LEGAL, ETHICAL, AND POLITICAL IMPLICATIONS

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 countervailing 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.

> ***"...begin a public discussion of the trade-offs between privacy, free speech, and democracy..."***

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. ■

REFERENCES AND NOTES

1. H. Allcott, M. Gentzkow, *J. Econ. Perspect.* **31**, 211 (2017).
2. N. Grinberg, K. Joseph, L. Friedland, B. Swire-Thompson, D. Lazer, *Science* **363**, 374 (2019).
3. K. H. Jamieson, *Cyberwar: How Russian Hackers and Trolls Helped Elect a President* (Oxford Univ. Press, 2018).
4. S. Vosoughi, D. Roy, S. Aral, *Science* **359**, 1146 (2018).
5. A. Friggeri, L. A. Adamic, D. Eckles, J. Cheng, in *Proceedings of the International Conference on Web and Social Media* (Association for the Advancement of Artificial Intelligence, 2014).
6. S. Aral, D. Walker, *Science* **337**, 337 (2012).
7. S. Aral, C. Nicolaides, *Nat. Commun.* **8**, 14753 (2017).
8. R. M. Bond *et al.*, *Nature* **489**, 295 (2012).
9. 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).
10. S. Messing, *Friends that Matter: How Social Transmission of Elite Discourse Shapes Political Knowledge, Attitudes, and Behavior*, Ph.D. thesis, Stanford University (2013).
11. J. L. Kalla, D. E. Broockman, *Am. Polit. Sci. Rev.* **112**, 148 (2018).
12. T. Rogers, D. Nickerson, Can inaccurate beliefs about incumbents be changed? And can reframing change votes? HKS Working Paper no. RWP13-018 (2013).
13. A. Peysakhovich, D. Eckles, Learning causal effects from many randomized experiments using regularized instrumental variables, in *Proceedings of the 2018 World Wide Web Conference* (International World Wide Web Conferences Steering Committee, 2018), pp. 699–707.
14. D. E. Broockman, D. P. Green, *Polit. Behav.* **36**, 263 (2014).
15. C. Dwork, Differential privacy: A survey of results, in *Proceedings of the International Conference on Theory and Applications of Models of Computation* (Springer, 2008).

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 *New York Times* in 2013 and visiting researcher at Microsoft in 2016. S.A. has received research funding from *The Boston Globe* and speaking fees from Microsoft. S.A. is an inventor on a related patent pending. D.E. has financial interest in Facebook, Amazon, Google, and Twitter. D.E. was a consultant at Microsoft in 2018. D.E. was an employee and consultant at Facebook from 2010 to 2017. D.E. has recently received funding from Amazon. D.E.'s attendance at conferences has recently been funded by DARPA, Microsoft, and Technology Crossover Ventures. D.E. is an inventor on a related patent, which is assigned to Facebook. S.A. and D.E. contributed equally to this work.

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