Evidence-Based Policy in Citizen Governance of Online Communities

Dissertation Proposal, MIT Media Arts and Sciences

J. Nathan Matías, B.A., M.A. (Cantab), M.S.

May 3, 2016

Dissertation Committee:
Ethan Zuckerman
Principal Research Scientist,
MIT Media Lab

Elizabeth Levy Paluck
Associate Professor
Department of Psychology Woodrow Wilson School
Princeton University

Tarleton Gillespie
Principal Researcher
Microsoft Research

Abstract

Critics of evaluating policy with field experiments have argued that experiments represent a turn away from democracy toward paternalism from experts. These randomized trials support inferences on the effects of policy ideas and the causes of their outcomes. In the fifty years since experiment-based policy evaluation was first proposed, they have remained the work of experts. Since then, the work of policymaking and the work of policy evaluation have expanded more widely through large-scale digital communications online. Online platforms now rely on hundreds of thousands of volunteers to set and enact "social policy" on what kind of social interactions are acceptable in their communities. Furthermore, randomized trials have also become common as methods for studying social behavior online, especially among platform designers. This dissertation offers five studies on the role of community-led randomized trials to support community policy-making in online communications and beyond.

The first study reviews the history of debates about the position and power of experimentation as a tool for democratic governance. This study will also discuss the potential and challenges of community-led field experiments to transform the policy work of online communities. The second study explores the idea of community policy in the social news platform reddit, where policies are debated, created, and enacted at scale by volunteers. This mixed-methods analysis uses trace data from over fifty thousand communities and interviews with moderators to observe how communities imagine, discuss, and advocate policy ideas.

The third and fourth studies are field experiments with communities on the social news site reddit. They test theories from social psychology while also evaluating community policies. The third study estimates the effect of posting community rules on newcomer behavior, expecting a trade-off between reductions in violations and reductions in participation. The fourth study compares the effects of normative and informational social influence on the politeness and impoliteness of large-scale online discussions.

The fifth study is a qualitative analysis of policy sense-making done by community members when they evaluate policies within their own communities using randomized trials.

Taken together, the research in this dissertation points to a future where greater access to experimental methods expands the capacity of citizen involvement in policy creation and evaluation

Introduction

Debates about the role of social experimentation in democratic societies have been present since the earliest discussions of evidence-based policy. Using randomized trials, researchers can draw inferences on the causes of policy outcomes and the effects of policy ideas [7, 21]. Because these field experiments require experts to conduct and interpret them, critics argue that policy experiments represent a turn away from deliberative democracy. In recent years, Thaler and Sunstein have proposed governance through "libertarian paternalism," where experimentally-derived knowledge is used to "steer people's choices in welfare-promoting directions without eliminating choice" [38]. Yet critics of libertarian paternalism have attacked it as a potentially-undemocratic form of social control, arguing that the assumptions and uses of experimental knowledge detract from citizen deliberation on policy [28].

Although the argument that expert evidence is undemocratic may be a recent re-reading of the Lippmann-Dewey debate [36], early proponents of empirical policy evaluation shared those concerns. In the 1960s and 70s, Campbell outlined an "experimenting society" where researchers would be "servants" of society working towards goals set by the public. Campbell argued that "even the conclusion drawing and the relative weighing of conflicting indicators, should be left up to the political process" [8].

Many of the limitations that kept experimental policymaking out of widespread public use have now been reduced through computational means [31]. Digital communications have expanded the capacity of citizens to coordinate, enabling them to take a direct role in the governance of common resources [14]. For example, in online discussion groups and social network platforms, hundreds of thousands and perhaps millions of people work as volunteer moderators, creating and enacting policies to govern social relations in their communities [19, 24]. In these digitally-mediated social contexts, data collection is commonplace, treatments are easy to deploy, and computer software

has broadened access to experiment design and analysis [29]. Employees at online platforms now commonly test the effects of design on social behavior [30, 3]. There is little keeping volunteer moderators from using similar methods to test their own policy ideas.

In this dissertation, I explore the potential and the challenges of community-led, field experiments to test moderation policies online. Volunteer moderators routinely create and enact policy on online safety, fairness, and conflict resolution in groups that range from a few dozen neighbors to tens of millions of subscribers. Communities commonly use software agents (social bots) to detect deviant behavior and automatically enact moderation policies [18]. Yet moderators rarely evaluate the effects of their governance work. In one case, volunteer moderators of the English language Wikipedia developed semi-automated software agents to respond to large-scale vandalism. Without the means to evaluate the effects of their governance systems, these moderators failed to notice, for several years, that their moderation practices had also created a dramatic, long-term decline in Wikipedia participation [25]. Had moderators tested the effects of their policies and adjusted them accordingly, participation in Wikipedia might not be as low. This dissertation introduces findings on the ways that volunteers make policy in online communities and the ways that they make sense of experiments to test the outcomes of those policies.

The research site for this dissertation is reddit, a social news platform with 244 million unique visitors, over 10,000 active communities, and over 150,000 moderator roles.[1] My prior work includes research on the volunteers who review reports of online harassment,[2] the nature of their governance work,[3] the information infrastructures they create,[4] and collective bargaining by these volunteer moderators with online platforms.[5] Building on ethnographic fieldwork and quantitative analysis of reddit moderation that I began in June 2015, this dissertation will offer mixed-methods findings on the nature of policy-making by volunteer moderators, quantitative findings from randomized trials co-designed with communities, and qualitative findings on the ways that communities negotiate experimental results in policy decisions.

This dissertation centers around the deployment of *CivilServant*, a software agent that facilitates the design, deployment, and monitoring of randomized trials in online communities.[6] CivilServant will be used for policy evaluation, supporting moderators and community participants to conduct field experiments to test their policy ideas. The system will offer communities a repertoire of dependent variables and study designs. It will coordinate randomized interventions at group, conversation, and individual levels. CivilServant will also manage some research ethics procedures such as consent, disclosure, and redress.[7]

CivilServant facilitates the vision outlined by Campbell in the 1960s, where the design and interpretation of experiments is open to a political process [8]. All research will be automatically opened for discussion by the community hosting the experiment. Results will be added to an open repository of citizen policy replications. These public conversations will form a key archive for qualitative findings on ways that communities negotiate experimental results about themselves from experiments that they have developed themselves.

Taken together, the chapters in this dissertation point to a future where greater access to experimental methods expands the capacity of citizen involvement in policy creation and evaluation.

[1] User count from April 2016. Moderator roles sampled in August 2015

[2] Matias, J. N., Johnson, A., Boesel, W. E., Keegan, B., Friedman, J., DeTar, C. (2015). Reporting, Reviewing, and Responding to Harassment on Twitter. Women Action and the Media.

[3] Matias, J. N. (2016) The Civic Labor of Online Moderators. Oxford Internet, Policy, and Politics Conference

[4] Matias, J. N. (2015, June 8) The Tragedy of the Digital Commons. The Atlantic. Retrieved from `http://www.theatlantic.com/technology/archive/2015/06/the-tragedy-of-the-digital-commons/395129/`

[5] Matias, J. N. (2016). Going Dark: Social Factors in Collective Action Against Platform Operators in the Reddit Blackout. CHI 2016.

[6] Although my methods involve developing a novel system, this dissertation does not include a systems paper.

[7] Since CivilServant supports participants to design and conduct their own experiments, it presents unique challenges for circumscribing ethical research practices. This research has been designed in consultation with the MIT Committee on the Use of Humans as Experimental Subjects (COUHES). The initial deployment of CivilServant will be limited to interventions and dependent variables that represent routine experiences and minimal risk. In any case where the online community requesting access to Civil Servant facilitates a market, mental health support, or discussion with a vulnerable population, that study will be independently submitted to the MIT COUHES. I will also exclude from participation any community that routinely organizes online to harm others. I also expect to host a subreddit for communities to reflect on and discuss the ethics of their own research.

1 Evidence-Based Policy in Online Communities

The volunteer moderators who develop and enforce social policies in online communities carry out a large share of policy work online [24]. A growing literature explores what it means for corporations to make social policy and test it, across ethical, regulatory, moral, and rhetorical dimensions [34, 10, 22, 32]. Yet even as corporations are creating and enforcing social policy to govern the behavior and speech of their users, these platforms also expect volunteers among their users to create and enact more localized policies. I argue that the contours of debates around policy experiments are very different when policy is being set, enacted, and tested by volunteers from the affected communities. This study considers the challenges and potential of randomized policy trials conducted by the people who are governed by social policies online.

This study will outline the history of debates over the role of evidence based policy, from Kurt Lewin's action research in the 1930s [2] and Cambell's experimental policymaking in the 1960s [7] to contemporary debates over libertarian paternalism [28] and corporate behavioral research [34]. Drawing from Stuart Geiger's theoretical work on the role of corporate researchers as civil servants of online governance [20], I will differentiate citizen experimentation from the prevailing model of top-down, corporate social research online [30]. The study will offer a framework for thinking about citizen control over the design and interpretation of experimental methods.

Research Status

I have already published six articles and workshop papers that incidentally address parts of this question,[8] but I have yet to combine them into a single argument.

2 Social Policy Creation and Enactment by Volunteer Moderators in Online Communities

Volunteer moderators within large-scale online groups develop sophisticated policy and governance regimes to guide participation towards cooperation and respond to problems [24, 17, 37, 6]. Across the web, a minimum of hundreds of thousands of volunteers participate in discussions of social policy and carry out actions aimed at upholding or enforcing those policies. This mixed methods study explores the work of policy creation and enactment by volunteer moderators of "subreddit" communities on the online platform reddit. This research sets out to explain how these moderators imagine, discuss, and share these policies with each other and with their communities.

Methods

To study policy creation, I will collect quantitative data from policy documents from the population of roughly 52,000 communities alongside moderation logs from a small sample of 6-12 communities. These logs contain every moderator action, from policy enactment to policy document editing.

I will also interview moderators, observe participation in moderator chat rooms, and analyze content from public discussions of policies with communities. I will focus my attention on moments of policy transition and tension.

[8] Matias, J. N. (2016). Participatory Field Experiments and Causal Inference For Monitoring and Advancing Social Justice in HCI. CHI 2016 Workshop Paper: Exploring Design, Social Justice, and HCI. San Jose, CA.
Keegan, B. C., Matias, J. N. (2015). Actually, It's About Ethics in Computational Social Science: A Multi-party Risk-Benefit Framework for Online Community Research. AAAI 2016 Symposium on Observational Studies of Social Media. Retrieved from http://arxiv.org/abs/1511.06578
Matias, J. N. (2016). Going Dark: Social Factors in Collective Action Against Platform Operators in the Reddit Blackout. CHI 2016. Retrieved from http://natematias.com/media/GoingDark-Matias-2016.pdf
Matias, J. N. (2015, June 8). The Tragedy of the Digital Commons. The Atlantic. Retrieved from http://www.theatlantic.com/technology/archive/2015/06/the-tragedy-of-the-digital-commons/395129/
Matias, J. N., Agapie, E., D'Ignazio, C., Graeff, E. (2014). Challenges for Personal Behavior Change Research on Information Diversity. Presented at the CHI 2014 Workshop on Personalizing Behavior Change Technologies.
Matias, J. N., Geiger, S. (2014). Defining, Designing, and Evaluating Civic Values in Human Computationand Collective Action Systems. In Second AAAI Conference on Human Computation and Crowdsourcing. Retrieved from http://www.aaai.org/ocs/index.php/HCOMP/HCOMP14/paper/view/9268

Research Status

I have been doing ongoing fieldwork with reddit moderators since June 2015. I currently have access to one full moderation log of a large subreddit. I have already written software for collecting information about policy making and enactment. I have not yet begun data collection.

3 The Newcomer Tradeoff Between Participation and Norm Compliance To Posted Rules in Large Online Communities

Subreddit communities on reddit often develop extensive policies governing the kinds of contributions they permit. In interviews, moderators report that users with deleted comments often send them apologies after their comments are removed, explaining that they didn't know the rules. To make those policies visible, moderators can add "sticky comments" to discussions, displaying rules next to the form where new comments are entered.[9]

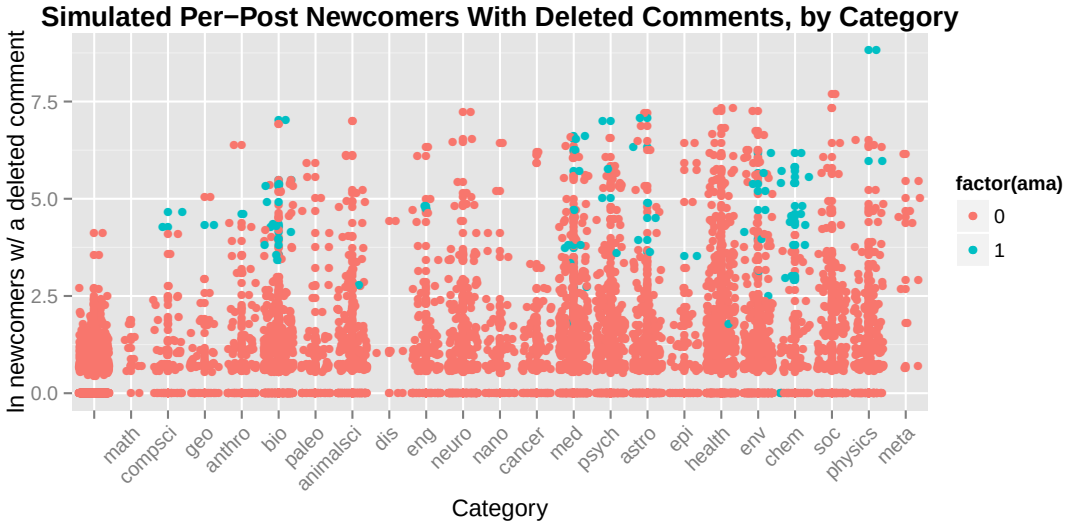
Does making users aware of rules through "sticky comments" have any effect on the behavior of newcomers? Field experiments in applied social psychology have tested effects of posting signs on littering behavior [35, 13], smoking in hotel rooms [12], environmental conservation by hotel guests [23], and crime reporting [4]. Yet posting rules might also have a negative effect on newcomer participation; commenters might not want to take the risk of having their comments removed.

This research tests the hypothesis that there is a trade-off between norm compliance and participation. Specifically, this experiment tests the effect of prominently posted "sticky comments" on the number of newcomers whose contributions that are removed by moderators, alongside its effect on the number of newcomer contributors.

Research Site

This research will be conducted on the social news platform reddit, within a subreddit community with 11 million subscribers. Moderators of this subreddit have granted me complete access to all subreddit data, including the moderation logs, and have agreed to conduct this study.

To support study design, participation data was collected from the sample of all public posts and comments that were not removed by moderators in the period from June 1, 2015 through September 1, 2015. To aid in the simulation of the dependent variables, moderation actions were collected from a 14 hour period on March 17, 2016.



[9] https://www.reddit.com/r/modnews/comments/1jr429/moderators_you_can_now_sticky_a_selfpost_to_the/

Methods

This study is a field experiment testing the effect of sticky comments on the number of newcomer deletions and the number of newcomer contributions [21]. The experiment has the following steps:

- In a three week period, I randomly assign certain days for all discussions receive "sticky comments" with subreddit policies
  - Control: no sticky comment
  - Treatment: sticky comment stating subreddit policies
- Moderators will receive a browser plugin that hides the sticky comments from them
- I observe per-post incidents rates of newcomers with at least one contribution that is removed by the community.
- I observe per-post incidents rates of newcomers.

Status of Research

Moderators of this community have already agreed to this experiment and are ready to begin as soon as IRB approval is given. A draft experiment pre-registration has been completed. Some development work is needed to monitor the dependent variables at scale over the experiment period.

4 The Effect of Positive Speech and Community Ratings on The Politeness and Impoliteness of Large Discussions Online

On reddit, popularity algorithms often flood large numbers of people into formerly-intimate community conversations. When a discussion is automatically highlighted on the front page, these unrelated newcomers know that they have been drawn together by a unique algorithmic decision, they don't see themselves as accountable to the subreddit's moderators, and they don't expect to see each other as a group again. At such moments, the scale of a conversation can overwhelm moderators while also weakening their usual means to prevent problems [33]. Moderators often hope to influence behavior by posting community rules. These posts establish subjective norms, beliefs about what important others think we should do that rely on a desire to be accepted [16]. However, that desire may not be relevant if commenters do not expect moderators' views to affect them.

Moderators and other reddit users have attempted two methods to facilitate polite conversations in cases where subjective norms may not apply. Firstly, they attempt to set a positive tone with early comments, hoping to set a pattern that others would follow. Secondly, subreddit participants sometimes upvote comments that are more positive and encouraging. Upvoters hope that those comments will become more prominent and that users seeking the reputation benefits from upvotes (all upvotes add to a public karma score for each user) will post similarly-positive comments when they see that those comments are upvoted. These practices correspond to theories on descriptive norms, where observations of the behavior of others offer another source of social norms [9, 1].

This experiment tests the effect of two interventions on the proportion of polite and impolite comments in discussions. The first intervention involves adding polite, encouraging comments to a new discussion thread. The second intervention involves adding polite, encouraging comments to a new discussion thread and then upvoting those comments prominently. The study tests the hypothesis that introducing early, polite comments will have a positive effect on the proportion of polite comments in a thread, and that the magnitude of the effect will be greater when those comments are upvoted.

Interventions will be carried out by community participants. Participating commenters will be notified several minutes in advance to prepare to add a positive comment, and then sent a link, via a chat interface, to the discussion in question. Up-voters will coordinated in a similar manner.

Even if interventions do have an effect within conversations, they might increase conflict or attract vandalism across the community. On reddit, users in conflict often accuse each other of

coordinated efforts to influence conversations. These accusations are sometimes used to justify vandalism of a community. In a qualitative followup analysis, I will explore the ways that the social context of an experiment shapes how the intervention and experimental results are received.

Research Site

The research site for this experiment is a high volume subreddit community with nearly 11 million subscribers and an average of 20,099 active daily participants. These participants make an average of 750 posts and 19,800 comments per day.[10] The subreddit has 28 moderators. When comments are coded with the Stanford Politeness Model (SPM) [11], an average of 21% of discussions, or 159 per day, have more than 10% of comments rated impolite. In contrast, 5% of discussions, or 41 discussions per day have more than 10% of their comments rated by the SPM as polite.[11] The group has an average of 6413 daily newcomers, who constitute 32% of daily commenters. 40% of impolite comments were from these newcomers and 25% of polite comments were from newcomers.

Quantitative Methods

This experiment tests the effect of early, polite comments on the proportion of polite and the proportion of impolite comments associated with an image post. The proportion of impolite and proportion of polite comments are related, separate dependent variables. A comment is polite or not; it is also impolite or not [26].[12] The experiment has the following steps:

- In a two-week period, we randomly draw one post within specified intervals of moderator availability [13] to receive positive, three-sentence comments within the first 2 minutes of the post's appearance:
  - Control: 3 neutral comments
  - Treatment A: 3 positive comments
  - Treatment B: 3 positive comments, each given 4-10 upvotes[14]
- Dependent Variables
  - proportion of *polite* comments sampled from each discussion, coded by subreddit participants
  - proportion of *polite* comments per discussion, scored by the SPM
  - proportion of *impolite* comments sampled from each discussion, coded by subreddit participants
  - proportion of *impolite* comments per discussion, scored by the SPM

Qualitative Methods

Qualitative data collection will include comments from conversation threads about the experiment after the results are announced. Qualitative analysis will focus on community reactions to the experiment and opinions about the acceptability of the tested intervention.

Research Status

The study design is still being refined. A large group of moderators and users are planning to coordinate on this study. This experiment will be pre-registered after feedback from participants.

[10] sample period: June 1 to August 31, 2015
[11] ratings were applied to up to 100 randomly sampled comments in a random sample of 1000 discussions
[12] the human coded dependent variable is still being refined
[13] the randomization method is still to be determined, based on what works best with moderators
[14] based on analysis of maximum comment upvotes among the 63,294 discussions sampled from Jun-Aug 2015

5 Evidence Based Policy in Online Communities

This study presents mixed-methods findings on the ways that communities make sense of experimental evidence in decisions about social policy in their online relations. In this study, communities are directly involved in the design, administration, and interpretation of randomized trials that occur within their community.

Methods

In this study, I use methods from participatory action research to collect evidence on the participation of communities in policy-evaluation technologies that they themselves guide and deploy [39, 2, 27]. I will focus my ethnographic research on community participation in the design and deployment of experiments conducted through the CivilServant system. Evidence collection is expected to follow the following process:

- *Initial design discussions about potential experiments with subreddit moderators and their communities:* These discussions yield qualitative evidence on the challenges of moderation and the parts of subreddit governance that moderators consider open to experimental methods. The reasons that communities decline to experiment will hold particular importance
- *Discussions refining and confirming experiment designs with subreddit moderators* will yield qualitative evidence on the challenges of experiment design with citizen groups
- *Community-wide discussions on the policy implications of experimental results* will yield qualitative evidence on the ways that participants make sense of experimental evidence in decisions about their own communities
- *Interviews with experiment subjects* will yield qualitative evidence on the ways that people who are part of experiments make sense of their participation in cases where those experiments are designed and conducted by their peers

Other sources of evidence will include quantitative data collected in the course of experiments, as well as ethnographic fieldnotes from this work of virtual ethnography [5, 15].

Research Status

The CivilServant system is in the early stages of its design and development. I have developed relationships and trust with reddit moderators over 11 months. Two subreddits have agreed to conduct the first experiments together, one more is discussing it, and one external researcher is already planning to use CivilServant for his own research when it is complete. These early experiments have established the design parameters for the CivilServant system.

I already have IRB approval for observational research of reddit trace data. I have had early conversations with the MIT Committee on the Use of Humans as Experimental Subjects and have determined an approach to research ethics that should acceptably manage the risks and benefits of citizen-led experimentation.

Timeline

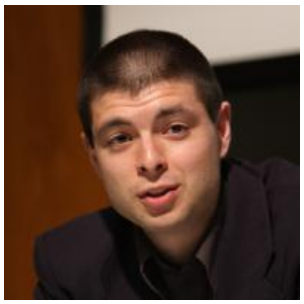
- *May 31, 2016*: Dissertation proposal critique
- *May 2016*: IRB process for Initial RCTs on reddit study
- *June 2016*: Initial RCTs on reddit begin
- *August 2016*: Initial RCTs on reddit complete
- *December 2016*: Pilot test of CivilServant, software for conducting RCTs on reddit
- *January - March 2017*: Deployment of CivilServant across multiple subreddits.

- *April - June 2017*: Qualitative analysis and writing for final Evidence-Based Policy study
- *July 2017*: Dissertation Defense

Resources Required

The resources required for this dissertation include substantial hardware resources and system administration for CivilServant, as well as systems for storing and processing qualitative and quantitative data. Some travel funds will be required for ethnographic fieldwork. Some of these travel costs will be covered by the Harvey Fellowship starting in September 2016.

Biography



J. Nathan Matias, B.A., M.A. (Cantab), M.S., is a PhD candidate at the MIT Media Lab Center for Civic Media with Ethan Zuckerman, a fellow at the Berkman Center for Internet and Society, and a two-time intern at Microsoft Research. Nathan has collaborated with a wide range of social media companies, news organizations, and advocates to study gender discrimination, harassment, and social movements online. He has authored 15 peer reviewed articles and workshop papers in computer science.

Nathan has worked in tech startups that have reached over a billion people and helped start a series of education and journalistic charities. He holds degrees in literature from the University of Cambridge and Elizabethtown College. A Davies-Jackson Scholarship recipient, Nathan has received the Nelson Award from the Association of Computer Machinery, the Riddick and Hugh Cannon Awards from the American Institute of Parliamentarians, and the Horton Fellowship award from MIT.

Dissertation Committee

Ethan Zuckerman, Principal Research Scientist, MIT Media Lab

Elizabeth Levy Paluck, Associate Professor, Department of Psychology, Woodrow Wilson School, Princeton University

Elizabeth Levy Paluck is an Associate Professor in the Department of Psychology and in the Woodrow Wilson School of Public and International Affairs at Princeton University. Her research is concerned with the reduction of prejudice and conflict, including ethnic and political conflict, youth conflict in schools, and violence against women. She uses large-scale field experiments to test interventions that target individuals' perceived norms and behavior about conflict and tolerance, including mass media and peer-to-peer interventions.

Tarleton Gillespie, Principal Researcher, Microsoft Research

Tarleton Gillespie is a Principal Researcher at Microsoft Research New England and an Adjunct Associate Professor in the Department of Communication and the Department of Information Science at Cornell University. His first book, Wired Shut: Copyright and the Shape of Digital Culture, was published by MIT Press in 2007. He is the co-editor (with Pablo Boczkowski and Kirsten Foot) of Media Technologies: Essays on Communication, Materiality, and Society (MIT, 2014). He is also the co-founder of the blog Culture Digitally. His current work considers the sociological implications of social media platforms and algorithms; his next book (Yale University Press, forthcoming 2017) examines how the governance of cultural values by social media platforms has broader implications for freedom of expression and the character of public discourse.

References

[1] Henk Aarts and Ap Dijksterhuis. The silence of the library: environment, situational norm, and social behavior. *Journal of personality and social psychology*, 84(1):18, 2003.

[2] Clem Adelman. Kurt Lewin and the origins of action research. *Educational action research*, 1(1):7–24, 1993.

[3] Eytan Bakshy, Dean Eckles, and Michael S. Bernstein. Designing and Deploying Online Field Experiments. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, pages 283–292, New York, NY, USA, 2014. ACM.

[4] Leonard Bickman and Susan K. Green. Situational Cues and Crime Reporting: Do Signs Make a Difference? *Journal of Applied Social Psychology*, 7:1–18, March 1977.

[5] Tom Boellstorff, Bonnie Nardi, Celia Pearce, and TL Taylor. *Ethnography and virtual worlds: A handbook of method*. Princeton University Press, 2012.

[6] Brian Butler, Lee Sproull, Sara Kiesler, and Robert Kraut. Community Effort in Online Groups: Who Does the Work and Why? *Human-Computer Interaction Institute*, January 2007.

[7] Donald T. Campbell. Reforms as experiments. *American psychologist*, 24(4):409, 1969.

[8] Donald T. Campbell. The Social Scientist as Methodological Servant of the Experimenting Society. *Policy Studies Journal*, 2(1):72–75, September 1973.

[9] Robert B. Cialdini, Carl A. Kallgren, and Raymond R. Reno. A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. *Advances in experimental social psychology*, 24(20):1–243, 1991.

[10] Danielle Keats Citron and Helen L. Norton. Intermediaries and Hate Speech: Fostering Digital Citizenship for Our Information Age. SSRN Scholarly Paper ID 1764004, Social Science Research Network, Rochester, NY, 2011.

[11] Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. A computational approach to politeness with application to social factors. *arXiv preprint arXiv:1306.6078*, 2013.

[12] Harold H. Dawley, John Morrison, and Sudie Carrol. The Effect of Differently Worded No-Smoking Signs on Smoking Behavior. *International Journal of the Addictions*, 16(8):1467–1471, January 1981.

[13] Yvonne AW De Kort, L. Teddy McCalley, and Cees JH Midden. Persuasive trash cans: Activation of littering norms by design. *Environment and Behavior*, 2008.

[14] Thomas Dietz, Elinor Ostrom, and Paul C. Stern. The struggle to govern the commons. *science*, 302(5652):1907–1912, 2003.

[15] Robert M. Emerson, Rachel I. Fretz, and Linda L. Shaw. *Writing ethnographic fieldnotes*. University of Chicago Press, 2011.

[16] Martin Fishbein and Icek Ajzen. Belief, attitudes, intention, and behavior. *An introduction to theory and research. Massachussets: Addison-Wesley*, 1975.

[17] Andrea Forte, Vanesa Larco, and Amy Bruckman. Decentralization in Wikipedia governance. *Journal of Management Information Systems*, 26(1):49–72, 2009.

[18] R. Stuart Geiger. Bots, bespoke, code and the materiality of software platforms. *Information, Communication & Society*, 17(3):342–356, 2014.

[19] R. Stuart Geiger and David Ribes. The work of sustaining order in wikipedia: the banning of a vandal. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work*, pages 117–126. ACM, 2010.

[20] Stuart Geiger. Does facebook have civil servants? On governmentality and computational social science. In *Workshop on Ethics for Studying Sociotechnical Systems in a Big Data World*, Vancouver, British Columbia, Canada, 2015.

[21] Alan S. Gerber and Donald P. Green. *Field experiments: Design, analysis, and interpretation*. WW Norton, 2012.

[22] Tarleton Gillespie. The politics of 'platforms'. *New Media & Society*, 12(3):347–364, 2010.

[23] Noah J. Goldstein, Robert B. Cialdini, and Vladas Griskevicius. A Room with a Viewpoint: Using Social Norms to Motivate Environmental Conservation in Hotels. *Journal of Consumer Research*, 35(3):472–482, October 2008.

[24] James Grimmelmann. The Virtues of Moderation. SSRN Scholarly Paper ID 2588493, Social Science Research Network, Rochester, NY, April 2015.

[25] Aaron Halfaker, R. Stuart Geiger, Jonathan T. Morgan, and John Riedl. The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist*, 2012.

[26] Claire Hardaker. *Trolling in computer-mediated communication: impoliteness, desception and manipulation online*. PhD thesis, Lancaster University, 2012.

[27] Gillian R. Hayes. The relationship of action research to human-computer interaction. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 18(3):15, 2011.

[28] Peter John, Graham Smith, and Gerry Stoker. Nudge nudge, think think: Two strategies for changing civic behaviour. *The Political Quarterly*, 80(3):361–370, 2009.

[29] Ron Kohavi, Roger Longbotham, Dan Sommerfield, and Randal M. Henne. Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery*, 18(1):140–181, 2009.

[30] Robert E. Kraut, Paul Resnick, Sara Kiesler, Moira Burke, Yan Chen, Niki Kittur, Joseph Konstan, Yuqing Ren, and John Riedl. *Building successful online communities: Evidence-based social design*. MIT Press, 2012.

[31] David Lazer, Alex Sandy Pentland, Lada Adamic, Sinan Aral, Albert Laszlo Barabasi, Devon Brewer, Nicholas Christakis, Noshir Contractor, James Fowler, Myron Gutmann, and others. Life in the network: the coming age of computational social science. *Science (New York, NY)*, 323(5915):721, 2009.

[32] Rebecca MacKinnon. Consent of the networked: The worldwide struggle for Internet freedom. *Politique étrangère*, 50(2), 2012.

[33] Adrienne Massanari. # Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society*, 2015.

[34] Michelle N. Meyer. Two Cheers for Corporate Experimentation: The A/B Illusion and the Virtues of Data-Driven Innovation. SSRN Scholarly Paper ID 2605132, Social Science Research Network, Rochester, NY, May 2015.

[35] Susan M. Reiter and William Samuel. Littering as a Function of Prior Litter and The Presence or Absence of Prohibitive Signs1. *Journal of Applied Social Psychology*, 10:45–55, 1980.

[36] Michael Schudson. The "Lippmann-Dewey Debate" and the Invention of Walter Lippmann as an Anti-Democrat 1985-1996. *International Journal of communication*, 2:12, 2008.

[37] Aaron Shaw and Benjamin M. Hill. Laboratories of Oligarchy? How the Iron Law Extends to Peer Production. *J Commun*, 64(2):215–238, April 2014.

[38] Cass R. Sunstein and Richard H. Thaler. Libertarian Paternalism Is Not an Oxymoron. *The University of Chicago Law Review*, 70(4):1159–1202, 2003.

[39] William Foote Ed Whyte. *Participatory action research.* Sage Publications, Inc, 1991.