GOVERNING HUMAN AND MACHINE BEHAVIOR IN AN EXPERIMENTING SOCIETY

J. Nathan Matias

B.A., Elizabethtown College (2005)
B.A. Cantab, M.A., University of Cambridge (2008)
M.S. Massachusetts Institute of Technology (2013)

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Media Arts and Sciences at the MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2017

Signature redacted

Author .....
Program in Media Arts and Sciences
April 19, 2017

Signature redacted

Certified by .....
Ethan Zuckerman
Associate Professor of the Practice
Program in Media Arts and Sciences
Thesis Supervisor

Signature redacted

Accepted by .....
Patricia Maes
Academic Head
Program in Media Arts and Sciences

Governing Human and Machine Behavior in an Experimenting Society

by J. Nathan Matias

Submitted to the Program in Media Arts and Sciences, School of Architecture and Planning, on May 18, 2017, in partial fulfillment of the requirements for the degree of Doctor of Philosophy in Media Arts and Sciences

Abstract

We live in a culture that depends on technologies to record our behavior and coordinate our actions with billions of other connected people. In this computational culture, humans and machines continue to perpetuate deep-seated injustices. Our abilities to observe and intervene in other people's lives also allow us to govern, forcing us to ask how to govern wisely and who should be responsible.

In this dissertation, I argue that to govern wisely, we need to remake large-scale social experiments to follow values of democracy. Using qualitative and quantitative methods, I spent time with hundreds of communities on the social news platform reddit and learned how they govern themselves. I designed CivilServant, novel experimentation software that communities have used to evaluate how they govern harassment and misinformation. Finally, I examined the uses of this evidence in community policy deliberation.

As we develop ways to govern behavior through technology platforms, we have an opportunity to ensure that that the benefits will be enjoyed, questioned, and validated widely in an open society. Despite common views of social experiments as scarce knowledge that consolidates the power of experts, I show how community experiments can scale policy evaluation and expand public influence on the governance of human and machine behavior.

Thesis Supervisor: Ethan Zuckerman
Title: Associate Professor of the Practice
Program in Media Arts and Sciences

This doctoral thesis has been examined by the following committee:

Signature redacted

Ethan Zuckerman

.....
Thesis Committee Chair
Associate Professor of the Practice
MIT Media Lab

Signature redacted

Tarleton Gillespie .

.....
Thesis Reader
Principal Researcher
Microsoft Research

Signature redacted

Elizabeth Levy Paluck

.....
Thesis Reader
Professor
Department of Psychology
Woodrow Wilson School
Princeton University

Twenty years after our argument about chemistry, my parents watched me defend my thesis, quite literally, when my MIT advisor Ethan Zuckerman slashed at me with a cosplay sword. Years of voracious intellectual wandering had led me unexpectedly to a community designed to help me flourish, and to one of the kindest, wisest, people that I have known.

Across our years together, Ethan showed me how scholarship should be focused on people as much as ideas. Studying with Ethan has been an extended masterclass in the facilitative leadership that makes everything he touches a joyful, collaborative, and inclusive endeavor. As Ethan shepherded my growth from a literary scholar and software engineer into an empirical researcher, he also modeled the kind of leader I want to be: where influence is based as much on amplifying and supporting others as sharing my own ideas.

I have found that to learn from Ethan is to grow together with the networks he so generously supports. Don't take him too seriously when he downplays his role. If Ethan's contributions seem small to him, it is only because he masterfully cultivates diverse, collaborative endeavors that beautifully exceed the sum of their parts. My dissertation would not have been possible without the years of practical experience in principled, public interest work that I gained with Ethan.

Although PhDs often end with a solitary endeavor, many hands move the project forward.

My chief supporter and encourager, Dr. H, has been kind, generous, and patient in uncountable ways, drawing from her own deep experience as an academic and as a supporter to thousands of other gradstudents. My remarkable parents also continue to encourage and inspire me.

One year ago, Merry Mou, an M.Eng. student at MIT, offered to work with me to make CivilServant a reality. As a collaborator on code and words, I have been delighted to see Merry grow on all fronts as she imagines new ways to apply her considerable technology abilities to questions of social value.

I am deeply grateful to my dissertation committee. Tarleton Gillespie has been a wonderful mentor since our time together at Microsoft. Betsy Paluck offered essential guidance in the early formation of the CivilServant project and its first experiments.

None of my PhD would have been possible without the effort invested by the r/worldnews and r/science communities, who spent many hours debating study designs and organized tens of thousands of people to participate in studies. Several people pored closely over my statistics before I shared experiment results with communities. Thanks, Feedmahfish and Martin Saveski! Several people also produced public datasets that became important to my dissertation, including Jason Baumgartner, Felipe Hoffa, Eric Gilbert, and Eshwar Chandrasekharan.

Great collaborations don't disappear when people graduate. This spring, the engineer and artist Sophie Diehl generously took a day to brainstorm novel procedural art based on data from CivilServant field experiments. Sophia Brueckner also offered helpful feedback. I'll blog soon about my first prototypes, which I briefly shared in the defense.

Over thirty scholars contributed their wisdom and their bibliographies to a literature review and a summit on online harassment. The collective knowledge we developed became an influential resource for many, including me.

I also had an amazing generals exam committee, who shepherded the hardest moments of my transition from purposeful designer to a question-asking social researcher. Benjamin Mako Hill is a force of nature who I am honored to have as a mentor, and Mary Gray has been a joyful, supportive, principled guide through discussions of research ethics and through my all-too-brief dips into qualitative research. As members of my MIT Master's thesis committee, Kate Crawford and Tom Steinberg inspired me to develop theoretically meaningful and genuinely impactful work that uses quantitative tools without being imprisoned by them.

The indescribably wonderful community at the Berkman Klein Center and the Cooperation Working group provided a long-term context for making that transition while still holding onto my passion for making a difference. BKC has provided an intellectual home and an important band of friends throughout my PhD years. Throughout our time facilitating the Cooperation Working Group, Brian Keegan was a wonderful collaborator and guide to the field of computational social science. Andrés Monroy-Hernández has been another influential role model at the Media Lab, the Berkman Klein Center, and then at Microsoft Research. Thanks for encouraging me to follow my dreams!

My six years at MIT have been funded by the Knight Foundation, the MIT Media Lab member companies, and the Harvey Fellowship. Thanks!

Several editors and publications worked closely with me throughout my PhD to develop articles and initiatives that have profoundly shaped my thinking. Spencer Kornhaber worked with me during the two years I facilitated @1book140 The Atlantic's Twitter book club, after Jeff Howe passed it into my hands. Adrienne LaFrance has been a wonderful collaborator and editor at the Atlantic as I developed my public voice. Becky Gardiner commissioned and edited a Guardian article that gave me confidence to follow through and apply intellectual history to contemporary issues in my dissertation.

I had so many good classes at MIT and Harvard. Betsy encouraged me to take an ICPSR class on field experiments with Donald Green and Costas Panagopoulos, which exposed me to valuable practices and ways of thinking in the art of experimentation. Thanks Don and Costas! On the Harvard Grad School of Education statistics train, I learned to value clear writing about statistics alongside valid methods. Thanks to Hadas Eidelman, Joe McIntyre, Shane Tutwiler, and Andrew Ho, for insisting on both! Sasha Costanza Chock introduced me to participatory research and why it matters; their influence runs deep throughout my work. I will always carry the inspiration of my first MIT class with Mitch Resnick, Sherry Turkle, and Karen Brennan, who made their class sparkle with their generous, respectful, and productive differences.

I have been lucky to meet many kindred spirits at MIT, and none greater than Ricarose Roque, Sayamindu Dasgupta, Erhardt Graeff, Rahul Bhargava, and the whole crew at the MIT Center for Civic Media. Thanks for being wonderful role models, dear friends, and thoughtful collaborators every step of the way, in play, reflection, and social action.

The amazing people of Women, Action, and the Media! set me on the path of prioritizing platform-independent research in our collaboration on harassment reporting. I am grateful to Amy Johnson, Whitney Erin Boesel, Brian Keegan, Jaclyn Friedman, Charlie DeTar, Jamia Wilson, and Adria Richards, for teaching me so much from that project.

I went on to study with Tom Leap at Elizabethtown College. Here's an image I photoshopped and put on his office door as an undergraduate.

Over the longer term, I am profoundly grateful to Tom Leap, who convinced me to study chemistry and eventually became my computer science professor during my undergraduate years at Elizabethtown College. I should also mention the late Tom Winpenny, who helped me find my voice for writing history, and whose undergrad class on technology and values first taught

me to seek the seeds of today's problems in the past. When I was an undergrad, my brother Jonathan invited me to participate in the American Institute for Parliamentarians, which sparked my early interest in community governance. I am also grateful to the Davies-Jackson Scholarship, Adrian Poole, Chris Warnes, Priyamvada Gopal, and the Union Society at Cambridge University, who helped me encounter important perspectives and supported my first stumbling attempts to reckon with questions of current affairs.

Many offered me a place to camp and work during my final year, including Matt Stempeck, Emma Pierson, Kaitlin Thaney, David Riordan, Marcus Gibson, Emily Gibson, James Docherty, Simon Berry, Wycliffe Hall Oxford, Matthew Jarvis, Blackfriars Oxford, John Lister, Diane Lister, Ugo Vallauri, Elizabeth Day, the Imperial College Faculty of Engineering, the Data and Society Institute, Jared Honeycutt, Wendy Quay, George Thampy, Amber Case, Kyle Drake, and the Boston Athenaeum. Other key resources and inspiration were provided by Lorrie LeJeune, Winter Moon Roots, Perry Hewitt, and Katherine Lo. Thanks everyone!

Contents

Chapter1

Introduction

We live in a culture that depends on technologies to record our behavior and coordinate our actions with billions of other connected people. In this computational culture, humans and machines continue to perpetuate deep-seated injustices. Our abilities to observe and intervene in other people's lives also allow us to govern, forcing us to ask how to govern wisely and who should be responsible. Two recent tragedies illustrate the need to ask these questions.

How Can We Govern Behavior Wisely?

When the French government first considered banning pro-anorexia websites in 2008, they were wrestling with the risks embodied by the integration of computational systems into everyday life (Carvajal, 2008). Internet publishing had made this potentially-harmful information easily accessible, and search algorithms were sharing it with an eager audience. As the French parliament and other governments continued to debate the idea for years, advocates pressured technology companies to limit the accessibility of media encouraging eating disorders.

In April 2012, the photo platform Instagram responded to public pressure by creating policies against content promoting suicide, self-harm, and eating disorders (Instagram, 2012). The company asked the public to report offending posts for removal and also redesigned their algorithms to prevent people from searching for self-harm images. As a platform, the company enabled anyone to publish images for free, generating revenue by collecting personal data and advertising to the people who use the service. A small platform with roughly 40 million users in 2012, Instagram's new policy affected 72% as many people as France's 55 million internet users.

Unfortunately, Instagram's policy may have caused the opposite outcome

from what they intended. Four years later, academic researchers re-examined their policy to find that communities that evaded Instagram's intervention received 15% more comments and 30% more likes after the ban (Chancellor, Pater, Clear, Gilbert, & De Choudhury, 2016). Other self-harm advocates may have simply moved to other platforms (boyd, Ryan, & Leavitt, 2011).

Because computational systems are used by people and automated systems monitor our behavior and intervene in the daily lives of billions of people, we have come to expect that they will play some role to govern our most difficult social problems. Yet like Instagram, we struggle to imagine wise governance because our debates about policy are not grounded in evidence. Governments have debated banning self-harm content for almost a decade, attempting to balance health risks with the risks from censorship. Years into those debates, the first evidence showed a strong possibility that there is nothing to balance. These policies, when put into practice, could actually increase health risks.

When Instagram first introduced its policies in 2012, the platform could have tested them with an experiment, developing strong evidence on the outcomes caused by banning self-harm image searches. By banning some terms and not others, perhaps for some users and not others, they could have learned if the ban achieved their goals on average. Similar tests are common on platforms, where A/B tests on advertising and sales routinely optimize company revenues (Kohavi, Longbotham, Sommerfield, & Henne, 2009). Without public evidence on the potentially-harmful effects of Instagram's policy in the lives of millions of people, the company persisted for years in governance practices with doubtful benefits.

I take the position that to govern behavior wisely, we need evidence on the outcomes of our attempts to govern billions of people in a computational culture. While policy researchers have taken similar positions since the 1960s (Oakley, 2000), experiments are rare in platform governance. Because platforms can host up to hundreds of social experiments per day (Kohavi et al., 2013), we should direct those methods to guide wise uses of the power that platforms provide.

Who Should Govern Behavior?

In April 2017, a New York Times Magazine article accused the ride-hailing company Uber of employing "hundreds of social scientists and data scientists" to

manipulate and exploit their drivers by "pulling psychological levers" (Scheiber, 2017). As a platform, Uber argued that its drivers were independent entrepreneurs who retained freedom over their driving choices. Yet researchers claim the company used social experiments to refine methods for influencing drivers to act against their own self-interest, reducing driver pay and increasing Uber's profits. (Rosenblat & Stark, 2016).

Because most platform experiments are done in secret by organizations that control the underlying software, these studies tend to concentrate power within platforms and away from the the people who use them, according to Rosenblat and Stark. In their nine-month ethnographic study with Uber drivers, they found that experiment-guided influence is fundamental to the company's business model. Social experiments empower the company to generate revenue from driver behaviors that favor the company but not the drivers themselves. Since drivers do not have access to research and data on their own behavior, the resulting "information assymetry" favors the company (Rosenblat & Stark, 2016).

Many leading scholars have argued for stronger professional ethics to protect the public from the risks of platform research (Grimmelmann, 2015; boyd, 2016; Zook et al., 2017). Yet ethics alone cannot solve the underlying imbalance of power. On one hand, social experiments can help society use platform power to govern wisely. On the other, platforms that are not accountable to the public hold most of this subtle power and the knowledge to use it. To gain the benefits of wise platform governance, we need to redesign the balance of power behind the research that will increasingly guide how billions of people are governed worldwide.

Community-Led Platform Policy Experiments

In the 1970s, the founding figure of policy evaluation Donald Campbell struggled with a similar dilemma of experimenting power. As he watched the U.S. government computerize data collection and centralize its social policies in an attempt to evaluate those policies more effectively, Campbell worried that the information asymmetry would weaken democracy. Campbell proposed an alternative thought experiment, a democratic "experimenting society" where local communities could evaluate their own policy ideas, share the results freely, and dispute decisions through deliberation and data alike. By spreading the

ability to design and interpret new studies, Campbell imagined that behavioral policymaking might more closely serve democratic values.

While Campbell's experimenting society remained a thought experiment, the structure of platforms makes community-led experiments a viable possibility for platform governance. Some platforms manage policymaking by delegating governance to the communities they host. Ever since the earliest platforms on the social internet, designers have delegated power to community leaders, sysops, and moderators, who create and enact policy in their communities (Rheingold, 1993; Bruckman, Curtis, Figallo, & Laurel, 1994; Butler, Sproull, Kiesler, & Kraut, 2002). Moderators tend to be accountable to the communities they govern, often collaborate with their communities on policy decisions, circulate policy ideas with each other, and sometimes advocate for their communities in platform-wide decisions. Furthermore, community moderators often possess the basic building blocks of policy experiments: access to behavioral data and wide abilities to intervene.

In this dissertation, I share early results of a project to grow our collective knowledge on the outcomes of platform policies while balancing the information asymmetries of platform power. Over the last two years, I have spent time with hundreds of communities on the social news platform reddit, learned how they govern themselves, designed novel software for platform-independent experiments, and supported two communities to evaluate policies governing human and machine behavior. I have also studied how communities circulate evidence, debate findings, and make decisions in conversation with policy experiments developed by communities themselves.

In this first chapter, I introduce the ideas and research findings in this dissertation.

In the second chapter, *Democratic Governance in an Experimenting Society*, I describe two risks we need to navigate as platforms take a greater role in governance. On one hand, if we intervene in millions of people's lives without evaluating the outcomes, we could make problems worse. On the other hand, if behavioral policymaking is done in secret, as it often is today, we face dangerous risks to human freedom. To move beyond this dilemma, I turn to the histories of factory management and policy evaluation.

In the 20th century, people with authority used behavioral research to guide corporate and government decisions at previously-unimaginable scales. These activities were resisted and reshaped by activists and researchers who worked to

reconcile experimentation with democratic values. Through the history of factory management, I retrace Valentine's proposals for group consent (Valentine, 1916) and Lewin's methods of action research (Lewin, 1944, 1946), where factory workers shape the design and interpretation of democratic management experiments. Following the history of policy evaluation, I introduce Popper's open society, where democratic citizens interpret findings from social experiments to reject ineffective or oppressive governance (Popper, 1947). I also summarize Campbell's proposal for an experimenting society, where civic participation includes participating in networks of local, democratic policy experiments (Campbell, 1998).

These historical examples remind us that research is design, and we can redesign our methods to follow democratic values. Rather than reject policy experiments or accept paternalism, we can develop ways to govern wisely with public participation and consent. I end this chapter by arguing for a democratic approach to policy evaluation on the platforms that govern our lives.

In the third chapter, *Governance by Volunteer Moderators Online,* I describe the work of volunteer moderators, examining the ways they are accountable to their communities, to platforms, and to other moderators. For over forty years, designers have created internet platforms to rely on volunteer moderators to govern community groups online. Because these moderators and their communities carry out their own policies and collect their own data within platforms, they possess the potential for community-led policy experiments. To describe this work, I spent time as an ethnographer with moderators on the reddit platform, a social news site that had over 148,000 moderator roles in July 2015. I show how moderators negotiate their relationships and their accountability with all three stakeholders in their everyday moderation work.

In the fourth chapter, *Community-Led Experiments in Platform Governance,* I introduce CivilServant, novel software I created that online communities on reddit use to evaluate their governance practices, share the results, and replicate other communities' policy experiments. I describe five central design considerations for any effort to develop community-led experiments: community participation, research ethics, experiment validity, transparency, and deliberative replication in an experimenting society. The CivilServant project addresses those design challenges with what I call the *community knowledge spiral,* a process for conducting governance experiments with community input and oversight (Figure 4-1). I also describe the software architecture that I developed

with Merry Mou to support this process and manage experiments. I evaluate the system by summarizing community responses to the experiments and reporting early uses of research findings by communities and the platform's designers. These early findings offer evidence that social experiments can generate informative governance knowledge in ways that are accountable to the people they govern.

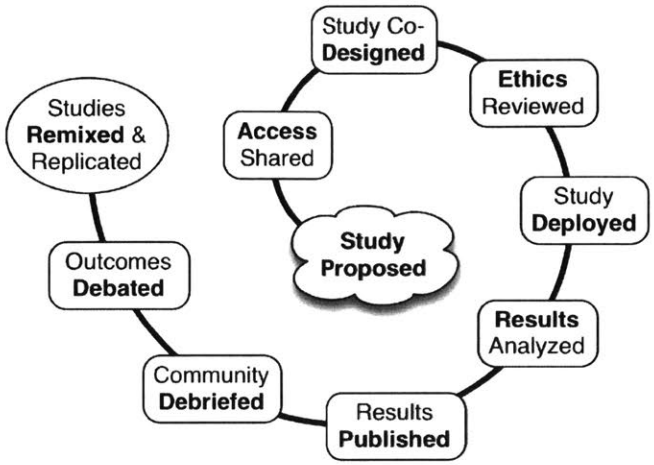


Figure 1-1: **The Community Knowledge Spiral**: CivilServant supports communities to test, replicate, and use knowledge from experiments that evaluate ideas for governing behavior through online platforms.

In the fifth chapter, *Preventing Online Harassment With Community-Led Policy Experiments,* I report results from a 14-million subscriber science discussion community that tested the effects of posting the rules at the top of discussions. In a policy experiment proposed and designed by community moderators that we conducted during August and September 2016, we randomly assigned rule messages to half of the community's 2,214 discussions and question-answer sessions. I found that posting the rules increased newcomer rule compliance by over 7 percentage points on average in the community, from 75% to 82%.

In the sixth chapter, *AI Nudges: Reducing the Algorithmic Promotion of Unreliable News by Influencing Social Behavior,* I report results from an experiment led by reddit's world news discussion group, which had 16 million subscribers in December 2016. Moderators in this community were concerned about the interactions between human behavior and reddit's algorithms that spread misleading and sensationalized tabloid news. They wished to intervene in ways that pro-socially influenced human and algorithm behavior while preserving

individual liberties. In this multi-armed experiment, possibly the first systematic effort to evaluate the prosocial effects of human influence on machine behavior, we encouraged commenters to fact-check unreliable news or fact-check and vote on the articles. Compared to no action at all, I found that both interventions increased the chance that individual commenters would link to further evidence in discussions. Surprisingly, I found that while encouraging fact-checking could reduce the promotion of unreliable news by reddit's popularity algorithms, I failed to find an effect from encouraging fact-checking and voting. As black box algorithms and AI systems play a greater role in human affairs, similar experiments may help us govern the unexpected interactions between human and machine behavior.

Community-led experiment results can only serve an open society if they are distributed, debated, and used. In the sixth chapter, *The Uses of Community Experiments in Online Policy*, I report the outcomes of adding experimental knowledge to the network of policy discussions and practices on reddit. I observe forces that shaped the ten-year history of a single policy adopted by reddit's politics community during the 2016 election. Drawing lessons from the field of *policy research utilization* started by Weiss in the 1970s (Weiss, 1979), I follow the spread and uses of evidence from two community-led experiments that I conducted with CivilServant.

Research evidence on reddit follows many of the patterns observed in government policymaking, yet the unique characteristics of platforms enable experiments to achieve rapid, widespread use in community policy. Since platforms are designed to propagate code and words, and since code and words remake the nature of those platforms (Kelty, 2005), new evidence and policy systems can spread widely across large numbers of communities. On reddit, evidence from community experiments appeared in policy conversations that were already underway when we published the results. Our findings informed the governance practices of individual moderators across the site and inspired community replications. The research also influenced over a hundred new communities when reddit designers tested new software that refined and replicated one of our experiments. I conclude this chapter with lessons from CivilServant for a society that uses experimental knowledge in behavioral policymaking.

In the epilogue, *Designing the Experimenting Society*, I reflect on the future of community-led policy experiments.

References

boyd, d. (2016, January). Untangling research and practice: What Facebook's "emotional contagion" study teaches us. *Research Ethics, 12*(1), 4–13. Retrieved 2017-04-13, from http://dx.doi.org/10.1177/1747016115583379 doi: 10.1177/1747016115583379

boyd, d., Ryan, J., & Leavitt, A. (2011). Pro-self-harm and the visibility of youth-generated problematic content. *ISJLP, 7*, 1. Retrieved 2017-04-13, from http://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/isjlpsoc7&section=4

Bruckman, A., Curtis, P., Figallo, C., & Laurel, B. (1994). Approaches to managing deviant behavior in virtual communities. In *CHI Conference Companion* (pp. 183–184).

Butler, B., Sproull, L., Kiesler, S., & Kraut, R. (2002). Community effort in online groups: Who does the work and why. *Leadership at a distance: Research in technologically supported work*, 171–194.

Campbell, D. T. (1998). The experimenting society. *The experimenting society: Essays in honor of Donald T. Campbell, 11*, 35.

Carvajal, D. (2008, April). French legislators approve law against Web sites encouraging anorexia and bulimia. *The New York Times*. Retrieved 2017-04-13, from http://www.nytimes.com/2008/04/15/world/europe/15iht-paris.4.12015888.html

Chancellor, S., Pater, J. A., Clear, T., Gilbert, E., & De Choudhury, M. (2016). # thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp. 1201–1213). ACM. Retrieved 2017-01-15, from http://dl.acm.org/citation.cfm?id=2819963

Grimmelmann, J. (2015). The Law and Ethics of Experiments on Social Media Users. Retrieved 2017-01-15, from http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2604168

Instagram. (2012, April). *Instagram's New Guidelines Against Self-Harm Images & Accounts.* Retrieved 2017-04-13, from http://blog.instagram.com/post/21454597658/instagrams-new-guidelines-against-self-harm

Kelty, C. (2005). Geeks, social imaginaries, and recursive publics. *Cultural Anthropology, 20*(2), 185–214. Retrieved 2017-04-06, from http://onlinelibrary.wiley.com/doi/10.1525/can.2005.20.2.185/full

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., & Pohlmann, N. (2013). Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1168–1176). ACM. Retrieved 2017-02-07, from http://dl.acm.org/citation.cfm?id=2488217

Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery, 18*(1), 140–181. Retrieved 2016-01-20, from http://link.springer.com/article/10.1007/s10618-008-0114-1

Lewin, K. (1944). The dynamics of group action. *Educational leadership, 1*(4), 195–200. Retrieved 2016-12-05, from http://www.ascd.com/ASCD/pdf/journals/ed_lead/el_194401_lewin.pdf

Lewin, K. (1946, November). Action Research and Minority Problems. *Journal of Social Issues, 2*(4), 34–46. Retrieved 2016-12-05, from http://onlinelibrary.wiley.com/doi/10.1111/j.1540-4560.1946.tb02295.x/abstract doi: 10.1111/j.1540-4560.1946.tb02295.x

Oakley, A. (2000). Experiments in knowing: Gender and method in the social sciences. Retrieved 2016-12-17, from http://philpapers.org/rec/OAKEIK

Popper, K. (1947). *The open society and its enemies.* Routledge.

Rheingold, H. (1993). *The virtual community: Homesteading on the electronic frontier.* MIT press.

Rosenblat, A., & Stark, L. (2016, July). *Algorithmic Labor and Information Asymmetries: A Case Study of Uber's Drivers* (SSRN Scholarly Paper No. ID 2686227). Rochester, NY: Social Science Research Network. Retrieved 2017-04-11, from https://papers.ssrn.com/abstract=2686227

Scheiber, N. (2017, April). How Uber Uses Psychological Tricks to Push Its Drivers' Buttons. *The New York Times*. Retrieved 2017-04-11, from https://www.nytimes.com/interactive/2017/04/02/technology/uber-drivers-psychological-tricks.html

Valentine, R. (1916, January). The progressive relation between efficiency and consent. *Bulletin of Taylor Society, 2*(1).

Weiss, C. H. (1979). The many meanings of research utilization. *Public administration review, 39*(5), 426–431. Retrieved 2017-03-17, from http://www.jstor.org/stable/3109916

Zook, M., Barocas, S., boyd, D., Crawford, K., Keller, E., Gangadharan, S. P., ... Pasquale, F. (2017, March). Ten simple rules for responsible big data research. *PLOS Computational Biology, 13*(3), e1005399. doi: 10.1371/journal.pcbi.1005399

Chapter2

Democratic Governance in an Experimenting Society

In this chapter, I describe two risks we need to navigate as platforms take a greater role in governance. On one hand, if we intervene in millions of people's lives without evaluating the outcomes, we could make problems worse. On the other hand, if behavioral policymaking is done in secret, as it often is today, we face dangerous risks to human freedom.

To move beyond this dilemma, I turn to the histories of factory management and policy evaluation. In the 20th century, people with authority used behavioral research to guide corporate and government decisions at previously-unimaginable scales. These activities were resisted and reshaped by activists and researchers who worked to reconcile experimentation with democratic values. These historical examples remind us that we can redesign our methods to follow our values. Rather than reject policy experiments or accept paternalism, we can develop ways to govern wisely with public participation and consent.

When the photo-sharing service Instagram banned hashtags promoting eating disorders in April 2012, it seemed like a victory for public health and civil society. Two years into its creation, the photo-sharing site was being used by peer support communities that encouraged anorexia and self harm as a "lifestyle choice." When journalists and a major UK charity drew public attention to the problem, Instagram took action, trying to disrupt these communities by making their common terms unsearchable. Unfortunately, the company's policies may have made the problem worse. According to a study published four years later, Instagram's actions drove the conversation underground and sometimes may have increased it. Communities that evaded Instagram's ban on search results received 15% more comments and 30% more likes after the ban (Chancellor, Pater, Clear, Gilbert, & De Choudhury, 2016). In the meantime, platforms have continued to apply similar policies to over a billion people worldwide in the hope of reducing terrorism (Lomas, 2017; Wark, 2016), hate speech (Citron & Norton, 2011), copyright violations (Seltzer, 2010), child pornography (Thakor, 2016), and human trafficking (Casteel, Thakor, Johnson, & others, 2011), without any publicly-accountable systematic evaluation of the outcomes.

Instagram could have evaluated their policy by testing the effect of banning certain searches. Yet behavioral experiments present their own risks to society, especially when experimenting power contributes to asymmetries of power. In 2017, a series of articles and class action lawsuits alleged that the digital ride-hailing platform Uber used secret behavioral experiments to systematically-influence drivers to act against their own self-interest in favor of company profits (Scheiber, 2017; Vaas, 2017). In a nine-month study of Uber driver experiences, Rosenblat and Stark concluded that platform algorithms and behavioral research allow Uber to control worker behavior while claiming that they offer their workers "freedom, flexibility, and entrepreneurship" (Rosenblat & Stark, 2016).

As Internet platforms become more pervasive and society turns to those platforms to govern social problems for billions of people, we need ways to gain the benefits of behavioral experiments while avoiding the risks of abuse that workers experienced with Uber. I argue that twenty-first century debates over the politics of large-scale online experiments resemble twentieth-century debates about the power of experiments in labor management and evidence-based policy. Across both debates, advocates and researchers who confronted

these dual risks created powerful ways to reconcile the benefits of social experiments with the values of democracy. We need to do the same.

I turn to the histories of labor management and policy evaluation because online platforms have combined these approaches into something new. Whether our online activities occur through for-profit platforms or nonprofits, Terranova argues that "the internet is about the extraction of value out of continuous, updateable work." Organizations that operate platforms develop indirect management practices to cultivate this value through freely-chosen actions (Terranova, 2000). To do so, platforms monitor their users' behavior, quantify it, include those measures in their reports to shareholders, and conduct large numbers of social experiments to maximize revenue from collective behavior (*First Quarter 2017 Results*, 2017). Platform operators also develop governance capacities to protect the value their users create (Prahalad & Ramaswamy, 2004) and to manage pressures from governments to regulate behavior (Gillespie, 2010). Behavioral researchers who work for platforms consequently combine management with governance, guiding productive behavior and evaluating policy goals. Just as management consultants on the early 20th century measured human behavior to manage labor and government social scientists in the 1960s computerized population-level data to manage policies on poverty and education, computational social scientists use data and the tools of experimentation to manage and govern our digitally-connected lives. Stuart Geiger argues that these data scientists are behavioral managers and policy evaluators at the same time, the corporate civil servants of the internet age (S. Geiger, 2015).

Like the data science of our time, management research and policy evaluation evolved in response to three converging factors: (a) expansions in the ability to structure and monitor behavior, (b) theories of experimentally-managed human endeavor, and (c) struggles to reconcile these developments with values of democracy. During the industrial revolution and the birth of evidence-based policy, debates about the politics of behavioral experiments produced powerful ideas about the role of research in democratic societies. These historical examples offer valuable ideas for redesigning the methods of experimentation to address social problems while preserving democratic values.

Autocratic and Democratic Factory Experiments

In 1944, as the social psychologist Kurt Lewin was setting out a vision for the role of social sciences in society, he worried that experimental research might be adopted in ways that would support autocratic, top-down governance of people's lives. For over a half-century, information technology and hierarchical theories of corporate management had co-evolved into what Lewin considered an autocratic system of organizing human cooperation, supported by experiments in labor efficiency. As a German Jew who fled Nazi Germany in 1933, Lewin argued that the management of factories and schools in the U.S. resembled the Nazi values he had left behind more than the vision of democracy he admired in the U.S., where he was now a citizen. Systematic managers had imposed measurements and experiments of worker production to empower autocratic management techniques. Lewin argued for democratic approaches to research and the management it supported.

Lewin responded with two fundamental insights on the politics of experiments in society, ideas that left deep influences on the growing fields of social psychology and organizational behavior. The first was a vision for the contribution that experimentation might make to democracy itself. In his 1944 book on group dynamics, Lewin argued that "it is essential that a democratic commonwealth and its educational system apply the rational procedures of scientific investigation also to its own process of group living" (Lewin, 1944, 120).

Lewin realized that this democratic vision could not be achieved through autocratically-imposed experiments; his second insight was a model for democratic participation in the experiment process itself. Lewin and his students' research in the 1940s transformed a long-standing argument over the assumed autocracy of research in industry, assumptions that persist today in the use of experiments online. For generations, firms and researchers had assumed that quantitative methods would lead to greater efficiencies than workers could imagine on their own. By including workers as designers in the experiment process, Lewin and his students offered quantitative evidence against these autocratic assumptions that many people considered scientific laws

Systematic Management

Large-scale data collection on human behavior became commonplace in U.S. industry during the hundred years before the second world war, argues Joanne Yates (Yates, 1993). As firms like railroads grew and became more geographically dispersed, they satisfied their need to coordinate with a growing toolset of information systems. Typewritten memos, forms, and graphs replaced handwritten letters. Telegraphs, telephones, and dictation machines replaced letterbags. As information piled up, firms developed filing systems to store and retrieve time-series information about activity, performance, and trends across a firm.

Measurement and coordination technologies developed in parallel with management philosophies that promised to help executives manage the growing complexities of industry. The philosophies of "systematic management" that developed from the railroad industry promoted two central ideas: (a) worker responsibilities should be standardized and defined by managers, and (b) information systems should monitor and evaluate the operation of the firm, including the adherence of workers to those manager-defined responsibilities (Yates, 1993, 10-11). Measurement of labor was applied to physical labor and also to what came known as white collar work (Mills, 1951), as management itself became systematized through standardized flows of responsibility, information technologies, and management theory (Yates, 1993, 11-15).

The most infamous system of measuring worker behavior was the "time and motion study," promoted by Frederick Taylor and other advocates of "scientific management." In this approach, engineers would use stopwatches to time repetitive actions towards a common task. Motion consultants would then work with management to set quotas and test more efficient work processes. Measurement for these studies was conducted by stopwatches or movie cameras with on-screen clocks (Nadworny, 1955). By the 1940s, companies had adopted a wider set of information technologies for ongoing labor monitoring, including forms, graphs, and ticket systems for tracking individual employees' completed tasks (Yates, 1993; Coch & French, 1948).

Theories of Behavior Under Systematic Management

Because systematic managers aimed to define and standardize the most efficient processes for achieving tasks across many workers, management experts

tended to assume that experts should design those processes. They believed that when workers used their own intelligence and agency to imagine factory processes, worker creativity put the system at risk. Measurement systems offered symbolic and rhetorical power to systematic managers, who positioned their decisions as scientific facts that could replace the inefficiencies arising from worker agency (Yates, 1993, 10-11).

Experts also tended to imagine tasks as independent units rather than connected, social activities. For example, time engineers would examine the number of distinct motions taken in bricklaying and test the efficiency and quality of alternative arrangements. The engineers would then mandate an efficient approach to that one part of bricklaying, alongside baseline task speeds that were used to set pay rates (Gilbreth, 1911). Advocates of Taylorism also held individual-centered beliefs about worker motivations, arguing that laborers tend to work as little as permissible when working in a group, a practice Taylor called "soldiering." To increase worker productivity, Taylor encouraged firms to adopt piece-work compensation models that paid workers based on how many units they completed above a baseline (Taylor, 1914). These personal incentives and punishments relied on standardized task rates that were initially set with time studies and carried out through the monitoring of workers in factory information systems.

Since advocates of systematic management held top-down, individual-centered views of behavior, management experts in the first two decades of the 20th century preferred what Lewin later called "autocratic" management. Systematic managers urged employers to negotiate directly with individual employees rather than allow them to organize as groups (Nadworny, 1955, 19-22)(Yates, 1993). Early systematic management consultants avoided working with unionized factories, and unions opposed piece-work arrangements, starting with the International Association of Machinists in 1903 (Nadworny, 1955, 25). Unions saw individually-calibrated pay rates, worker monitoring, and increased production as opportunities for firms to treat workers unfairly and cut jobs. Taylorists worried that negotiating with workers as a group would create inefficiencies in systems that they believed were based on scientific fact (Nadworny, 1955, 49-51).

Despite these fundamental disagreements, the resulting unrest between workers and firms forced greater communication and coordination between them. This coordination took three forms in the period from 1915 to 1929: corporate

welfare practices, arguments for applying social psychology to democratic management, and "joint research" between unions and firms. Yet none of these developments succeeded at changing widespread assumptions that worker agency and group decisions were inefficient.

The corporate welfare movement offered an alternative to unions, personalizing the management relationship according to Yates. Shop conferences and representative shop committees discussed problems and proposed changes to work processes. Internal company magazines featured employee perspectives and communicated norms to employees. Firms also offered benefits and funded local projects including libraries, clubhouses, and city beautification. Yet according to Yates, these widespread efforts were not primarily seen as ways to improve to production. Managers adopted corporate welfare practices defensively, to discourage union organizing and other forms of collective employee power (Yates, 1993, 17-18).

Arguments for collective worker agency were rejected in the early years of scientific management. Writing in the Bulletin of the Taylor Society in 1916, Robert Valentine argued that achieving efficiency required "organized consent as well as individual consent" from workers. Valentine served as an external investigator at the Watertown Arsenal after a 1911 strike against scientific management. Having participated in the 1912 U.S. Commission on Industrial Relations, Valentine had been seeking ways to reconcile the methods of scientific management with the collective interests of workers (Nadworny, 1955). To convince other scientific managers to consider collective forces in labor management, Valentine published diagrams of the social forces at play in factory decision-making (Figure 2-1). He urged the Taylor Society to learn from the new field of "social psychology" to develop a research agenda on the roles of democracy and consent in efficient production (Valentine, 1916). The Bulletin published ten strongly unfavorable rebuttals along Valentine's article urging worker consent. Critics argued that Valentine under-valued the efficiency benefits of scientific management, that all rational people would naturally consent to the "purity" of scientific knowledge, that trade unions were wasteful, selfish endeavors, and that workers could not be trusted to understand the complexities of business. Valentine's suggestions were not pursued by the society (Nadworny, 1955; Gomberg, 1985).

Throughout the 1920s, unions and advocates of systematic management found common ground through joint Bureaus of Standards that negotiated

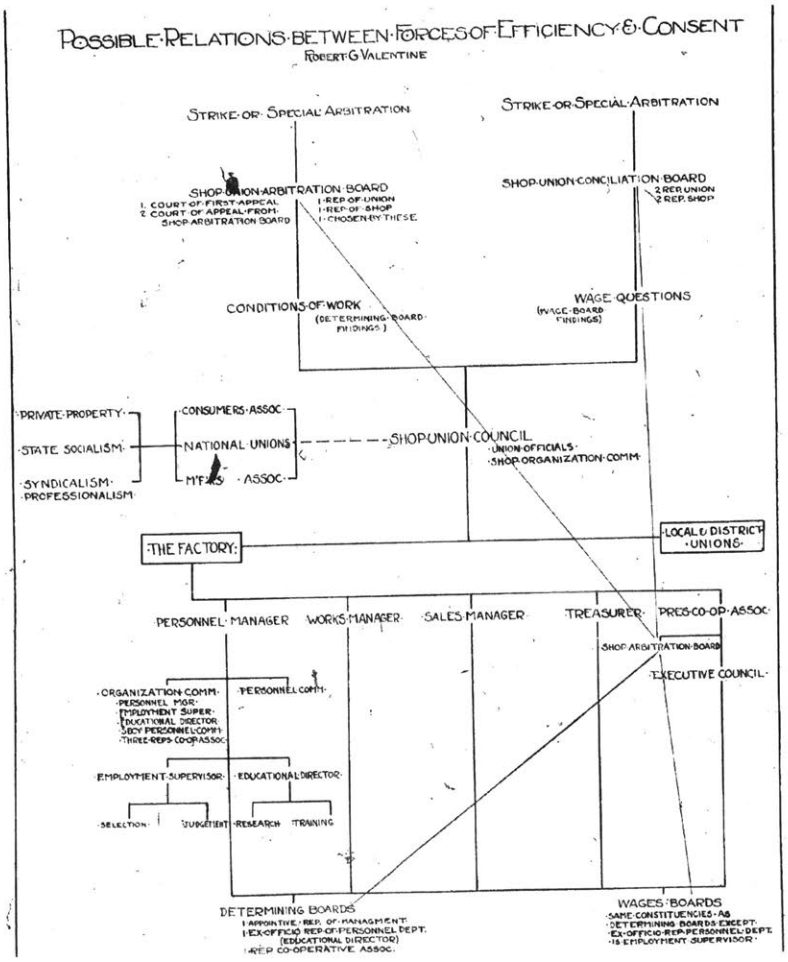


Figure 2-1: In 1916, Robert Valentine mapped factory social structures to argue for organized consent in factory research

over the baseline goals set through time studies. By 1929, several unions developed their own engineering teams, who participated in negotiations over the details of systematic management policies. In 1929, the United Textile Workers of America(UTW) and the Naumkeag Steam Cotton Company of Salem, MA set up a program of "joint research" managed by both the union and the company. Decisions about changes in production were made through collective bargaining. By 1932, the conditions of the Great Depression were pressuring the company to cut costs, and the findings of joint research led the union to agreeneed. In response, workers left the UTW entirely and went on strike. Cut out of the dispute, the union dropped its joint research pilot program at Naumkeag (Nadworny, 1955, 122-41).

While managers and workers in U.S. industries found new ways to communicate and coordinate throughout the 1920s and 30s, many still saw them as opposing interests. Isolated voices advocated that worker collaboration with management could achieve greater productivity. Yet projects including Valentine's proposal of organized consent and the Naumkeag joint research initiative were rare. Instead, employers adopted a corporate welfare approach to defend their companies against the risk that workers might organize and develop meaningful power.

Action Research at Harwood Manufactoring

By 1939 when Kurt Lewin's student Alex Bavelas moved to Marion, Virginia as an embedded psychologist at the Harwood Manufacturing Co, worker participation seemed to be a necessary inefficiency in factory management. Six years earlier, New Deal policies including the National Industrial Recovery Act (1933) and the National Labor Relations Act (1935) required good-faith collective bargaining between companies and unions (*National Industrial Recovery Act of 1933*, 1933; *National Labor Relations Act of 1935*, 1935). Harwood was led by a new president and chairman, Alfred Marrow, who had finished his psychology PhD at NYU two years earlier (Marrow, 1969; "Dr. Alfred Marrow, A Psychologist 72", 1978). The recently-established Marion plant was Harwood's first in the U.S. South, and Marrow saw an opportunity to improve factory operations through experimental psychology. The resulting studies inspired new thinking on the relationship between democratic practices and experimenting power.

At the Harwood pajama plant, Marrow had hoped to show that corporate

welfare philosophies and scientific management could coexist productively. The plant, which opened with roughly 300 primary school educated employees, carried out corporate welfare practices including healthcare, worker orientations, shop conferences with workers, occasional company-wide votes, and recreation programs (Coch & French, 1948; Burnes, 2007). Among scientific management practices, the company adopted the piecework employment model. A time engineer measured the efficiency of each task to set upper and lower target rates for work. The company then paid workers differentially, based on tickets that employees were given when they completed a work unit. Workers who failed to meet the plant minimum were fired after a probationary period. Employees reportedly hoarded tickets to meet the baseline quotas on days when they felt unwell (Coch & French, 1948).

Throughout the 1940s, Harwood faced challenges common in Taylorist firms, struggling with "grievances about the piece rates that went with the new methods, high turnover, very low efficiency... and marked aggression against management." When work practices were changed and tasks were re-timed by managers, even experienced workers would quit after failing to regain former levels of efficiency. The company responded with layoffs in negotiation with the union, which "did little or nothing to overcome the resistance to change" (Coch & French, 1948). Alfred Marrow saw the company as unusually engaged with workers, but the workers clearly did not. Neither group seemed confident that manager and worker expectations were compatible.

Alex Bavelas and Lewin's other students reportedly overcame this stalemate by creating a new paradigm of research practice, one where workers played an active, sometimes democratic role in the design of factory research.[1] Since the company was already collecting data on rates of production and turnover alongside qualitative data on worker "aggression" toward managers, these measures became the dependent variables of their studies. Although none of the Harwood experiments would be recognizable as valid randomized trials today, they did compare means and standard deviations of these outcomes between treatment and control groups (Coch & French, 1948).

In one of his first publicly-reported studies, Bavelas tested the effect of worker discussions and votes on piece-work targets. While the majority of the firm received targets from managers and time engineers, a second, treat-

[1] All records of these studies in this article come from researchers and from Marrow himself. Without their voices, it is difficult to assess worker views of these studies.

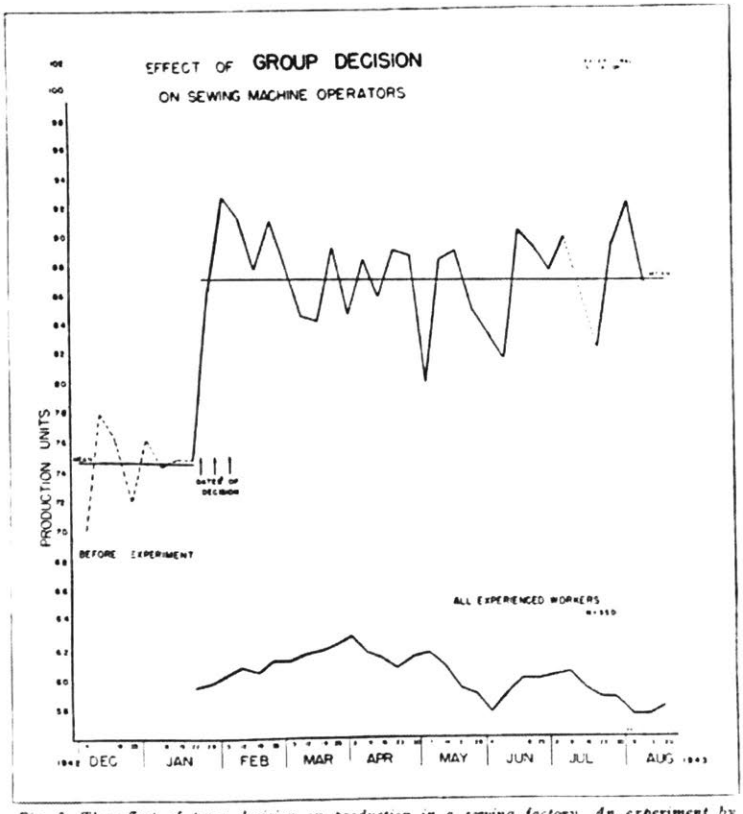


Fig. 2. The effect of team decision on production in a sewing factory. An experiment by Alex Bavelas shows a marked permanent rise in production after decision. As comparison, the production level of experienced workers is given during the same months.

Figure 2-2: At Harwood Manufacturing, Alex Bavelas charted productivity to observe the changes associated with group decisions (Lewin, 1944)

ment group was asked to discuss and decide on group goals. Contrary to a generation of management theory, workers who set targets as a group increased production (Figure 2-2), a finding that Bavelas confirmed in further studies (Marrow, 1969, 144). The actual interventions in these early studies were poorly documented—Lewin describes the intervention as a "team decision," while Marrow, writing 30 years later, describes a setting where individual workers decided on personal targets in the context of a group discussion. Both of them describe a series of studies where Bavelas tested a wide range of variations on worker group agency in the research design: group discussions, individual decisions within groups, group votes, and personal goal-setting. Reportedly, all of these interventions offered some increase in worker productivity compared to what Lewin called autocratic management techniques (Lewin, 1944, 197)(Marrow, 1969, 144-45).

When Lester Coch and John French replaced Bavelas at Harwood in the mid-40s, they formalized these questions into more clearly defined experiments. They designed the studies to contribute to psychological theory while also answering practical questions for Harwood. Observing that high output workers failed to recover former levels of productivity after moving to a new task, they explained this outcome with a theory that connected the physical, individual, managerial, and group forces that tend to reduce or increase a person's production. By conducting experiments during attempts to change those forces, Coch and French were able to study practical factory questions in ways that could also evaluate their theories. In their findings, Coch and French argued that group resistance to work tasks and piece rates constituted a major limitation on productivity and employee retention. They tested the effect of using elected representatives to work with the time and motion engineer to set group baselines, as well as a participatory process for designing tasks (Coch & French, 1948). In one case study, Coch and Marrow supported supervisors to do their own research and make a group decision. They found that supervisors let go of their false stereotypes about the productivity of elderly women only after discussing their individual findings together a group (Marrow & French, 1945). Across the Harwood studies, Coch and French argued that group resistance could be reduced by including workers in designing tasks, setting rates, and conducting their own research. These reductions in resistance to change, they argued, enabled large increases in productivity (Coch & French, 1948).

Experimenting Power in Factory Management

Time studies and other experiments in systematic management met organizational needs at a time when U.S. businesses were growing dramatically in size and complexity. Yet these information and research technologies co-evolved with theories of management that viewed collective worker agency as an obstacle to business goals. Researchers were consequently unable to imagine or test the benefits of democratic work processes. Instead, many scientific managers reacted to collective action as a risk to their expertise and the very idea of experimentation itself.

Kurt Lewin and his students re-appropriated workplace information systems to prioritize worker ideas and group decisions in factory research. Keeping the Hardwood factory's measurement systems in place, they redesigned the experiment process to test the effects of group decisions compared to autocratic ones. Their experiments brought benefits to everyone involved. Workers stayed longer with the company. Harwood improved its efficiency and its ability to adapt to changing conditions. Lewin and his proteges developed and tested new theories on the social forces beyond individuals that shape human behavior, making fundamental contributions to social psychology. Lewin and his colleagues drew from those findings to develop a vision for citizen involvement in the design and interpretation of social experiments in democratic societies. Writing during the second world war, several years before his death at age 57, Lewin argued that "Efficient democracy means organization, but it means organization and leadership on different principles than autocracy" (Lewin, 1944). By placing the component parts of experiment design and analysis under democratic control, Lewin's research demonstrated how researchers could re-imagination along those democratic principles.

Authoritarian and Democratic Policy Evaluation

Twenty years after Kurt Lewin's death, another experimental psychologist, Donald Campbell, was wrestling with concerns about the authoritarian or democratic role of experiments in society. By 1971, the U.S. government had begun investing in large-scale electronic data collection to administer and evaluate national policies on housing, poverty, and education. Campbell's writings on social experiments and policy evaluation had become assigned reading in uni-

versities and government offices (Campbell, 1981). Yet as experiments become more common in the US, Campbell asked if a democratic, open society could ever be an experimenting society (Campbell, 1998). Campbell was responding to the philosopher Karl Popper's proposal that evidence-based "social engineering" in open societies could offer a democratic alternative to authoritarian rule. But as U.S. involvement in the Vietnam war continued to escalate and his student's sentiments turned leftward, Campbell speculated that growth in evidence-based policy could lead instead to a state governed by experimenters (Campbell, 1981, 482). In a series of grant applications, talks, and essays Campbell imagined alternative democratic cultures of policy experimentation. While many of his proposals seemed impractical to contemporaries, Campbell's idea of a democratic "experimenting society" has gained new relevance in a time of plentiful behavioral experiments online.

Social Engineering in the Open Society

Campbell's effort to reconcile evidence-based policy with democracy draws from Karl Popper's book, *The Open Society and Its Enemies.* Writing as an Austrian exile in New Zealand during the second world war, Popper describes two kinds of governance: open and closed societies. In closed societies, authoritarians govern and manipulate the public towards utopian goals on the paternalistic principle that "the learned should rule" (Popper, 1947, 107). In open societies, the public is encouraged to evaluate and criticize government decisions "so that bad or incompetent rulers can be prevented from doing too much damage" (Popper, 1947, 107).

Popper calls policy researchers in authoritarian societies "utopian social engineers." These researchers introduce and evaluate social policies with a predetermined goal toward an idealized society. The agenda for this "ultimate political aim" is set by the authoritarian state or by revolutionaries who wish to dramatically remake society (Popper, 1947, 138). Popper uses the history of eugenics in Western philosophy as his primary example of utopian social engineering. He argues that the disruptive, authoritarian experiments of utopian eugenicists tend to override public objections and consent:

> The reconstruction of society is a big undertaking which must cause considerable inconvenience to many, and for a considerable span of time. Accordingly, the Utopian engineer will have to be

> deaf to many complaints; in fact, it will be part of his business to suppress unreasonable objections. But with it, he must invariably suppress reasonable criticism also. (Popper, 1947, 140-41)

Popper contrasts this authoritarian, closed, utopian research with the democratic "piecemeal social engineering" of an open society. Piecemeal policy evaluators, he argues, should "adopt the method of searching for, and fighting against, the greatest and most urgent evils of society, rather than searching for, and fighting for, its greatest ultimate good" (Popper, 1947, 139-40). Popper hoped that long-term improvements could be achieved through many small policy evaluations considered through a democratic process. Every small adjustment to economic policy, the justice system, and social change could be evaluated experimentally, debated, and discarded if they failed to achieve their stated goals or were rejected democratically.

In Popper's comparison of utopian and piecemeal engineers, he argues that social experiments can be conducted in authoritarian or democratic ways. To limit the risks of authoritarianism, Popper makes a bundled argument in favor of incremental, democratic policy evaluation. Arguing against Soviet and Nazi approaches to social engineering, Popper favors small policy experiments over large changes such as revolutions or state-sponsored eugenics. According to Popper, smaller policies are less disruptive, more easily understood, and more easily challenged by the public. He argues that this piecemeal approach leads to a kind of steady social evolution away from injustice, guided through democratic processes rather than violence or oppression. But Popper was not himself a methodologist or policymaker. His key insight about experiments in The Open Society is not about the supposed relationship between the size of an experiment and its compatibility with democracy. Rather, Popper contributed to the political philosophy of experimentation by observing that the design of a social experiment has a direct relationship with its democratic or authoritarian potential.

Popper also worries that social engineers would come to govern directly. He compares these experimental sovereigns with Plato's hope for a dictatorship led by philosopher kings:

> education has a definite political function. It stamps the rulers and it establishes a barrier between the rulers and the ruled (This has remained a major function of higher education down to our own

time.) Platonic wisdom is acquired largely for the sake of establishing a permanent political class rule. (Popper, 1947, 130)

Responding to Popper in the late 1960s and early 70s, Donald Campbell draws on Popper's concerns when he outlines an "experimenting society" where social scientists are "methodological servants" of democracies rather than makers of policy (Campbell, 1973). Over the previous decade, Campbell had observed the growing pains of U.S. government data processing and evidence based policy. Even as Campbell advocated for more widespread, more rigorous policy evaluation, he wished to avoid Popper's dystopian vision of a society governed by social scientists.

The Great Society and the U.S. War on Poverty

The U.S. government first began systematically evaluating its social policies after President Johnson announced a War on Poverty in his 1964 State of the Union Address. Johnson's vision closely resembled Popper's piecemeal social engineering: declare a deprivation to be reduced and organize a long-term effort across successive policies to achieve that goal. Johnson also introduced systematic evaluation into the U.S. government. During the past two decades, the government had employed social scientists to analyze wartime policies and military operations. Many researchers, including Kurt Lewin and Donald Campbell, spent the second world war conducting social research for the military. Starting with Kennedy's presidency, Secretary of Defense Robert McNamara, former CEO of Ford Motor Company, had extended systematic evaluation across the U.S. military. In 1965, Johnson ordered all federal departments to adopt similar policy evaluation methods. Like the systematic management of industry, policy evaluation co-evolved with new information technologies and the theories of management that made use of them (Oakley, 2000, 201) (Rossi & Wright, 1984; Jardini, 2000).

According to historian David Jardini, the systematic management of U.S. social programs was imported from efforts to manage the U.S. military with lessons learned from U.S. manufacturing (Jardini, 2000, 318). In 1961, McNamara began expecting military leaders to offer quantitative justifications of weapons development and military strategy. After Kennedy's death, when Johnson declared a "war on poverty," he was implying that efforts to end poverty would be pursued with similarly-systematic and quantitative efforts (Jardini,

2000, 327). Citing private letters exchanged during the drafting of War on Poverty legislation, the historian Jardini argues that moves towards policy evaluation paralleled the shift in industry away from worker agency. As McNamara's proteges moved into social policy, they argued for systematic policy evaluation out of a distrust that local communities could develop effective poverty reduction interventions with federal resources (Jardini, 2000, 334).

Many policymakers believed that advances in computer technology created new possibilities to centrally manage social problems, according to the historian Andrew Meade McGee (McGee, 2007). Rhetoric about the transformative capabilities of large-scale electronic data processing bolstered beliefs in the possibility of achieving Great Society goals like eliminating poverty. By 1966, president Johnson issued a "Memorandum on the Use and Management of Computers by Federal Agencies," urging departments to adopt computers. Noting that the federal government already employed 2,600 computer systems, he argued that automated data processing could help the U.S. government "administer the huge and complex social security and medicare programs" to "provide better service to the public" (Johnson, 1966). At the time, the Social Security Administration had dramatically expanded the role of electronic computers to deliver the administration's Medicare healthcare program. The agency was given one year to enroll over 19 million Americans into the program, compute ongoing medicare payments and monitor civil rights compliance by healthcare providers across the country (Puckett, 2010).

Even as U.S. federal agencies used new computational capacities to administer programs, they were also required to use the data to evaluate the outcomes of those programs (Williams, 1971; Oakley, 2000). Yet the demand for evaluation outpaced the ability of these "analytical offices". According to Walter Williams, chief of the Research and Plans Division in the U.S. Office of Economic Opportunity, U.S. government policy evaluation was an aspiration rather than a genuine capacity in the 1960s. Social scientists were not experienced at asking the questions posed by policymakers, they worked too slowly, and they had little influence on data collection. Furthermore, researchers predominantly relied on correlation-based inferences that were subsequently dismissed in policy debates. Williams concluded that the social sciences were "unlikely to produce a consistent flow of studies of major relevance to social policymaking" (Williams, 1971, xiv).

Campbell was drawn into these evaluation circles after analysts became in-

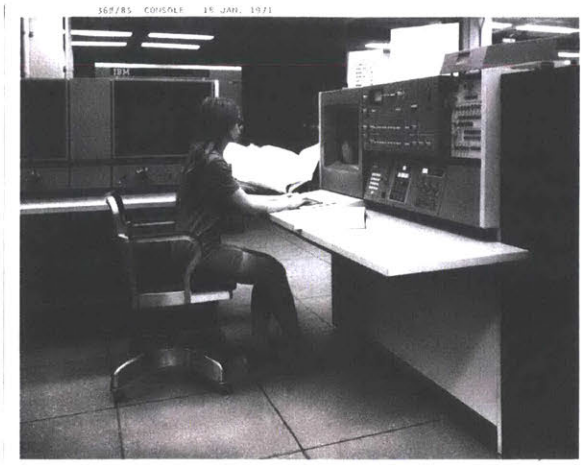


Figure 2-3: The IBM System/360, an early family of interoperable computers launched in 1964, offered a common computing platform usable by the U.S. military, intelligence services, NASA, and social services (agent, 1971)

spired by a book on research methods he co-authored in 1963, "Experimental and Quasi-Experimental Designs for Research" (Campbell & Stanley, 1963; Shadish, Cook, & Leviton, 1991). For the first time, the U.S. government had the will to systematically evaluate the outcomes of its policies. Campbell and Stanley's book offered the means (Oakley, 2000, 202-3).

In 1970, as the U.S. government transitioned into Nixon's presidency, Great Society analysts advocated for changes in the research methods, implementation process, and politics of evaluation.[2] Campbell, Williams, and others argued for causal evaluations that could give policymakers guidance on the effects of social programs (Campbell, 1969) (Williams, 1971, 7-8). Furthermore, many of the early War on Poverty programs were designed in Washington without careful implementation planning. Williams and other later commentators argued that pilot tests and community involvement might have prevented implementation problems that resulted from this top-down approach (Williams, 1971, 11) (Oakley, 2000, 236). But Williams also suggested a further change. He encouraged analysts against neutrality. Instead, he urged social scientists to engage fully in agency politics: proposing policy ideas, monitoring

[2]In the subsequent decade, policy experiments expanded dramatically. One bibliography of studies grew from 83 entries to 245 between 1974 and 1978 (Oakley, 1998). But in 1971, when Campbell first asked if the open society could be an experimenting society, causal policy research was still rare. Many of the subsequently-famous randomized trials on tax policy, housing, preschool, and criminal justice were still underway or still being planned.

implementation, and advocating for their view of the best government actions (Williams, 1971, 9).

Experiments In an Open Society

Campbell may have hoped that the idea of an experimenting society could reconcile decreasing trust in government with growing optimism about the benefits of experimental methods. Even as policymakers adopted experimental methods in the 1960s, Campbell's students and colleagues were advocating for civil rights and questioning the role of the U.S. government in Vietnam and Latin America. Looking back decades later, Campbell recalled factors that offered a "guiding impact" on his idea for an experimenting society: the leftward shift of students, a seminar with urban planning students, discussions with Czechoslovakian policy researchers, and conversations with a Chilean psychologist who had worked under the Marxist Allende government (Campbell, 1998, 1981).

In his 1971 speech "The Experimenting Society," Cambpell builds on Popper's distinction between authoritarian and democratic experimentation to imagine a vision for social experiments in an open society. In this imagined society, participation in experiments would be a basic a part of our civic lives like voting, reading the newspaper, or showing up to a town council meeting:

> Participation in policy experiments is more akin to participating in democratic political decision making than to participating in the psychology laboratory (Campbell, 1998).

According to Campbell, this imagined society would prioritize policy evaluation: political ideas would be judged by "action research" that tested outcomes (Campbell, 1998, 38), with a "willingness to change once-advanced theories in the face of experimental and other evidence" (Campbell, 1998, 41). Yet in this "utopian" thought experiment, Campbell suggests that the means of policy experimentation be broadly distributed throughout society to maximize diversity, autonomy, consent, and deliberation.

Citizens would set the goals and design of experiments in an experimenting society, Campbell argues. Rather than solely testing proposals from social scientists, he urges methodologists to find ways to include "individual participation and consent at all decision levels possible" (Campbell, 1998, 42).

Non-experts should hold substantial power when interpreting experiment results, Campbell argues. By "legitimating and facilitating evaluation by non-professional participants and observers," researchers might overcome over-simplifications assumed by many quantitative measurement of social problems. Arguing that "those who have situation-specific information... make the best critics," he suggests that an experimenting society should "provide these non-professional observers with the self-confidence and opportunity to publicly disagree with the conclusions of the professional applied social scientists" (Campbell, 1998, 55).

Citizens should be supported to critique statistical results, according to Cambell, who argued that "citizens not a part of the governmental bureaucracy will have the means to communicate with their fellow citizens disagreements with official analyses and to propose alternative experiments." To support this citizen policy creativity, Campbell urges governments to publish experiment results and support "recounts, audits, reanalyses, and reinterpretations of results" by non-governmental groups (Campbell, 1998, 42).

Overall, Campbell imagines experimentation as a basic building-block of democratic participation. Arguing against "the enforcement of assigned treatments," he encourages researchers to craft reliable study procedures that treat people as "co-agents directing their own society" rather than "passive recipients" (Campbell, 1998, 49).

Community Replications in an Experimenting Society

In "The Experimenting Society," Campbell also suggests ways that locally-developed experiments could combine across communities to develop collective knowledge. Citizen-led social experiments could generate large numbers of regional policy results, Campbell points out. As these results are made public, the combined data could cross-validate shared knowledge on the potential outcomes of a policy in different regions. Federal researchers could then evaluate the most promising policy ideas that bubble up through community replications (Campbell, 1998, 42).

To introduce this idea, Campbell describes the statistical limitations of top-down evaluation. Describing an imaginary experiment mandated by the U.S. Congress, he points out that the results of any single study involve substantial uncertainty and can be doubted even when conducted well. Campbell argues

that "the dependability of reports... comes from a social process rather than dependence on the honesty and competence of any single researcher" (Campbell, 1998, 51). The social process he proposes, "contagious cross-validation," supports local communities to conduct new experiments and question the findings of research by other communities.

Imagine, for example, that several towns wish to reduce traffic accidents. In Campbell's experimenting society, those towns might design and evaluate their policy ideas. Participants in local communities could debate the experiment designs as "adversarial stake-holders" (Campbell, 1998, 53). Car drivers, freight drivers, cyclists, and push-cart street vendors in this hypothetical example might need to negotiate competing priorities to develop a policy to test (Merrill, 1964). A nearby city might also consider the policy and develop a parallel experiment that matched urban considerations. These early evaluators might be followed by other communities eager to test and apply the policies themselves, attracting the attention of a state or federal government that might support more widespread evaluation. "After five years, we might have 100 locally interpretable experiments," writes Campbell, imagining the benefits of combining results to assess the overall impact of a policy idea (Campbell, 1998, 52). If the data from each experiment were made widely available, the statistics and interpretation of each experiment could be evaluated by independent researchers and constituencies with the "mutual criticism" of contesting views (Campbell, 1998, 51).

To prioritize the validity and influence of individual studies, Campbell's contemporaries had urged government researchers toward greater power. Campbell re-imagined social processes of experimentation to offer a democratic alternative. In "The Experimenting Society," he encourages researchers to remake experimental methods and the surrounding social processes to expand community participation. He argues that including citizens as "co-creators" can improve the internal implementation of a study. Campbell also describes the statistical advantages of supporting many local, democratic policy evaluations. Campbell argues that a democratic society that conducts plentiful local experiments from multiple standpoints can validate findings more thoroughly than centralized experimentation.

Campbell was responding to a period of dramatic growth in the goals of U.S. social policy. New, computerized approaches to social research were co-evolving with centralized theories of population-level governance that relied

on those information systems. Campbell drew from Popper's Open Society to recognize the risks to democracy of this emerging form of social engineering. As a methodologist, he was also able to design alternatives to the authority-driven approaches to social governance imported from U.S. industry and military. With the Experimenting Society, Campbell charted out a rough politics of citizen participation in the design and discourse of social experiments, an approach whose statistical advantages would be difficult for authoritarian methods to replicate.

Experiments and Platform Power

In the early 21st century, digital technologies now mediate a substantial proportion of human activity. The platforms that carry out this work collect intimate data on the everyday actions of billions of people, ushering in what some researchers have called a new age of computational social science (Lazer et al., 2009). Growing fields in the social and computer sciences have evolved to support platforms to manage behavior at scale (Kraut et al., 2012). Just as early 20th century factory data collection co-evolved with industrial management theories and 20th century government data collection developed alongside theories of social reform, the era of computational social science is co-evolving with theories for governing human behavior through the behavioral sciences.

Management and governance power on the internet is primarily held by platforms, organizations whose software digitally mediate human interactions by publishing, sorting, and routing information for individuals and groups. As communications systems or markets that mediate information, platforms would face legal liabilities in the U.S. if they took direct responsibility for people's behaviors. Yet platform value models depend on their ability to coordinate large-scale behavioral patterns in productive directions. When positioned as markets, platforms promote entrepreneurship and freedom while harnessing collective labor to generate revenue (Prahalad & Ramaswamy, 2004; Terranova, 2000). When positioned as communications systems, they generate advertising revenue by shaping attention patterns, promoting individual liberty to disclaim responsibility for harmful speech and behavior (Gillespie, 2010; Citron & Norton, 2011).

Because platforms position themselves as impartial brokers who support free expression and entrepreneurship, they tend to use indirect approaches to

governance and management. Market platforms like the ride-hailing system Uber craft incentive structures and interface designs that preserve individual agency while influencing workers to behave in ways that benefit the company, on average (Rosenblat & Stark, 2016; Scheiber, 2017). Communications platforms like Instagram take a similar approach to governance where possible. By increasing the difficulty of searching for images encouraging self-harm, Instagram's designers likely believed they could influence potentially-harmful forms of attention without restricting speech rights (Chancellor et al., 2016).

The art of influencing human behavior beneficially at scale while avoiding coercion is central to the idea of "libertarian paternalism," a political philosophy promoted by Thaler and Sunstein. In their view, behavioral scientists and policymakers can achieve substantial public goods through small changes to the "choice architecture" of everyday life, which they call "nudges" (Thaler & Sunstein, 2003). For example, a government could increase its revenues without raising taxes by changing a default setting in motor vehicle registration form so that citizens need to explicitly opt out of donating funds to the government. In this scenario, no one has been forced to pay, but the government can still be guaranteed a revenue increase (Thaler, Sunstein, & Balz, 2014).

Platform designers are the choice architects for billions of people's daily behavior, and platforms routinely use experiments to evaluate and adjust platform architectures in real-time. Developing a platform into a highly adaptable social machine takes many years of labor from hundreds of software engineers. Yet once achieved, platforms can deploy hundreds of changes per day, testing the average outcomes of those changes through specialized systems for conducting field experiments (Bakshy, Eckles, & Bernstein, 2014). By 2013, researchers at Microsoft estimated that they were conducting up to 300 experiments a day with users of the Bing search engine (Figure 2-4) (Kohavi et al., 2013). In 2017, researchers at AirBNB were reporting over 400 experiments per week (Parks, 2017). On many platforms, high demand for behavioral decision-making has led companies to implement systems that automatically test alternatives and make adaptive decisions in real-time about the optimal choice architecture to guide behaviors beneficial to the company (White, 2012).

Because platforms are designed to manage behavioral experiments across large proportions of humanity, society stands to benefit greatly from the ability to evaluate platform policies. It is incomprehensibly tragic when platforms like Instagram try to address serious societal risks without using their substan-

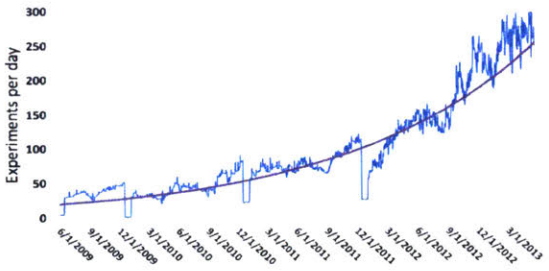


Figure 2-4: By March 2013, Microsoft Was Conducting Up To 300 Field Experiments Per Day On Bing.com (Kohavi et al., 2009)

tial experimental apparatuses to evaluate the effect of their efforts. When platforms do evaluate their policies, they rarely publish the results.

Despite the potential benefits of social governance through platforms, large-scale secret corporate behavioral research contributes to power asymmetries that platforms sometimes abuse. In the case of Uber, designers and behavioral scientists crafted choice architectures that caused many of their so-called independent contractors to take consistent actions that undermine their self-interest in favor of the company (Rosenblat & Stark, 2016). Like Popper's utopian engineers, these researchers ignored the complaints of their workers and customers, allegedly automating mass deception to influence behavior and maximize profits (Vaas, 2017). As the public continues to pressure platforms to take a greater role to govern society, Geiger has argued platform researchers are becoming a new kind of civil servant—digital policymakers and evaluators who remain largely unaccountable to the billions of people they govern (S. Geiger, 2015).

Democratic Experiments for Platform Governance

In the years since Campbell's proposal for an experimenting society, communications platforms, computer software, and education have reduced many of the difficulties of establishing an experimenting society. Interventions and data collection have become dramatically less expensive and easier to deploy, leading to an environment where experiments are already plentiful. For an experimenting society to be realized, communities, software engineers, and researchers would need to take advantage of these advances to re-design how experiments are conducted.

Research designers should develop methods for supporting communities to develop their own policy evaluations online. On many parts of the internet, communities already manage their own local policies and create novel policy systems where platforms do not explicitly delegate that governance to them (R. S. Geiger, 2016; Butler, Sproull, Kiesler, & Kraut, 2002). Even in settings of digital labor where workers are purposefully dis-empowered, workers have found ways to adapt a software environment to collect data for governing the behavior of those who offer them work (R. S. Geiger, 2014). Wherever communities manage their own policies, an experimenting society might support those communities to evaluate their policy work. Many systems already exist to support non-experts to design marketing and sales experiments; lessons from those systems can be adapted to community-led experiments.

Software systems and computer science education have already broadened data literacy in the social sciences. People often encounter quantitative representations of their personal and social lives through technology platforms, which often offer resources for analyzing that data. On one popular platform where children create stories and animation, researchers supported children as young as 11 years old to conduct their own analyses of their social behavior (Dasgupta, 2016). While people in an experimenting society will continue to need the support of the data advocates described by Campbell, additional work on data literacy and non-expert data analysis systems can broaden the pool of those advocates.

Campbell argued that an experimenting society would need expanded capacities to share, re-analyze, and dispute the results of experiments. Since Campbell's time, the developments that have enabled platform governance have also created the conditions for the disputatious network of local policy evaluators that Campbell imagined. Open source and open data infrastructures rapidly distribute datasets and the means to analyze them. Since many policies are encoded partly in software, platforms often provide the conditions to distribute and to disagree with policy ideas, as well as evaluate them.

If platform governance were conducted in an open, experimenting society, communities would be able to propose, design, and reject policy evaluations. They would borrow policy and experiment designs from each other in relationships of cooperation and criticism. Individual participants could influence policy evaluations, develop their own interpretations, and make arguments for what a community would decide.

Even though new experiments would contribute to a growing pool of evidence, the primary indicator of a flourishing experimenting society would not be the body of knowledge it produced. Instead, the value developed by this network of disputatious, experimenting communities would be visible in the creativity of community policy ideas, the tailoring of policies to local contexts, the criticism and rejection of evaluated policies, and the compromises that communities develop together. In Popper's closed society, social engineers choose the best option based on available data. Research findings in an open, experimenting society would serve as just one resource in wider public deliberations about platform governance. In such a society, the power of any single study would be smaller and more diffuse, while research would be more plentiful and widespread.

Platform Governance in an Experimenting Society

Online platforms have created the capacity to monitor and influence the behavior of billions of people, and they now face increasing demands to address fundamentally-important social problems. Experiments can provide valuable methods for testing the outcomes of platform governance and protecting the public from harmful platform policies. Unfortunately, emerging approaches to computational social science risk tilting the balance of power away from a consenting, democratic public and into the care of platform researchers who mostly work in secret. If used wisely by an open society, theories of libertarian paternalism could achieve collective goods while preserving individual liberties. Yet autocratic, unaccountable choice architectures could also preserve an appearance of liberty while significantly exploiting human life on average—a subtle kind of platform oppression.

In the early twentieth century United States, systematic management theories co-evolved with information technologies and experimentation practices to manage human production at previously-unimaginable scales. In the 1960s and 70s, growing computational capacities and research methods increased government confidence in large-scale efforts to reduce social ills. At both times, people with power came to believe that their goals could only be achieved by further consolidating their power, avoiding the inefficient or methodologically-messy disputes of the governed.

In each period, struggles to reconcile experimenting power with demo-

cratic values generated powerful new ideas for the role of experimentation in society. During the industrial transition to systematic management, advocates and researchers including Kurt Lewin developed models where workers collectively decided the design and evaluation of their own labor. Later, as government data collection and social policies expanded, Karl Popper and David Campbell developed political theories and research methods for non-authoritarian policy evaluation of social reforms.

Taken together, their ideas remind us that research is design, and we can redesign our methods to follow democratic values. At Harwood, Lewin and his students introduced democratically-determined interventions: workers discussed and voted on changes in their labor that would be tested experimentally. Campbell, Popper, and Lewin all made contributions to the idea that those affected by experimentally-evaluated power should have an influential voice in the uses of that power. Popper argued that an open society should learn learn from experiment results to approve or abolish policies through democratic means. Lewin and Campbell realized that workers and citizens would need to deliberate with statistics as well as words in such a society. Finally, Cambpell combined his methodological expertise with Popper's political theory to imagine constellations of communities and advocates, all using the tools of statistics to critique, construct, and cross-validate social reforms in a democratic society.

As we develop the foundational patterns of platform governance in the 21st century, we have a fresh opportunity to ensure that that the benefits will be enjoyed by an open society. While experimental methods can help us avoid harmful uses of that power, they could also enable new kinds of platform oppression. By re-designing the methods of experimentation, we may yet make progress on urgent social problems while maintaining democratic values.

References

agent, U. U. g. (1971, January). *National Security Agency IBM 360/85 console.* Retrieved 2017-01-26, from `https://commons.wikimedia.org/wiki/File:Supercomputer_NSA-IBM360_85.jpg`

Bakshy, E., Eckles, D., & Bernstein, M. S. (2014). Designing and Deploying Online Field Experiments. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 283–292). New York, NY, USA: ACM. Retrieved 2015-

10-03, from http://doi.acm.org/10.1145/2566486.2567967 doi: 10.1145/2566486.2567967

Burnes, B. (2007). Kurt Lewin and the Harwood Studies The Foundations of OD. *The Journal of Applied Behavioral Science, 43*(2), 213–231. Retrieved 2016-12-03, from http://jab.sagepub.com/content/43/2/213.short

Butler, B., Sproull, L., Kiesler, S., & Kraut, R. (2002). Community effort in online groups: Who does the work and why. *Leadership at a distance: Research in technologically supported work*, 171–194.

Campbell, D. T. (1969). Reforms as experiments. *American psychologist, 24*(4), 409. Retrieved 2016-04-21, from http://psycnet.apa.org/journals/amp/24/4/409/

Campbell, D. T. (1973, September). The Social Scientist as Methodological Servant of the Experimenting Society*. *Policy Studies Journal, 2*(1), 72–75. Retrieved 2016-04-23, from http://onlinelibrary.wiley.com.libproxy.mit.edu/doi/10.1111/j.1541-0072.1973.tb00128.x/abstract doi: 10.1111/j.1541-0072.1973.tb00128.x

Campbell, D. T. (1981). Comment: Another perspective on a scholarly career. *Scientific inquiry and the social sciences*, 453–501.

Campbell, D. T. (1998). The experimenting society. *The experimenting society: Essays in honor of Donald T. Campbell, 11*, 35.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and Quasi-Experimental Designs for Research* (1st ed.). Wadsworth Publishing.

Casteel, H., Thakor, M., Johnson, R., & others. (2011). Human Trafficking and Technology: A framework for understanding the role of technology in the commercial sexual exploitation of children in the US. Retrieved 2017-03-28, from http://www.iu.edu/~traffick/_resources/_literature/_research/_assets/Human-Trafficking-and-Technology.pdf

Chancellor, S., Pater, J. A., Clear, T., Gilbert, E., & De Choudhury, M. (2016). # thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp. 1201–1213).

ACM. Retrieved 2017-01-15, from http://dl.acm.org/citation.cfm?id=2819963

Citron, D. K., & Norton, H. L. (2011). Intermediaries and hate speech: Fostering digital citizenship for our information age. *Boston University Law Review, 91*, 1435. Retrieved 2017-01-15, from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1764004

Coch, L., & French, J. (1948). Overcoming resistance to change. *Human Relations, 1*, 512–532. doi: 10.1177/001872674800100408

Dasgupta, S. (2016). *Children as data scientists : explorations in creating, thinking, and learning with data* (Thesis, Massachusetts Institute of Technology). Retrieved 2017-05-12, from http://dspace.mit.edu/handle/1721.1/107580

Dr. Alfred Marrow, A Psychologist 72. (1978, March). *The New York Times*. Retrieved 2016-12-16, from http://www.nytimes.com/1978/03/04/archives/dr-alfred-marrow-a-psychologist-72-an-expert-on-group-dynamics-he.html?_r=0

First Quarter 2017 Results (Tech. Rep.). (2017, May). Facebook, Inc. Retrieved 2017-05-13, from https://investor.fb.com/investor-news/press-release-details/2017/Facebook-Reports-First-Quarter-2017-Results/

Geiger, R. S. (2014). Successor Systems: The Role Of Reflexive Algorithms In Enacting Ideological Critique. *Selected Papers of Internet Research, 4*. Retrieved 2016-04-23, from http://spir.aoir.org/index.php/spir/article/view/942

Geiger, R. S. (2016, June). Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society, 19*(6), 787–803. Retrieved 2016-08-29, from http://dx.doi.org/10.1080/1369118X.2016.1153700 doi: 10.1080/1369118X.2016.1153700

Geiger, S. (2015). Does facebook have civil servants? On governmentality and computational social science. In *Workshop on Ethics for Studying Sociotech-*

nical Systems in a Big Data World. Vancouver, British Columbia, Canada. Retrieved from https://cscwethics2015.files.wordpress.com/2015/02/geiger.pdf

Gilbreth, F. B. (1911). *Motion study: A method for increasing the efficiency of the workman.* D. Van Nostrand Company.

Gillespie, T. (2010). The politics of 'platforms'. *New Media & Society, 12*(3), 347–364. Retrieved 2017-01-17, from http://nms.sagepub.com/content/12/3/347.short

Gomberg, W. (1985). The Historical Roots of the Democratic Challenge to Authoritarian Management. *Human Resource Management, 24*(3), 253–269. Retrieved 2016-12-14, from http://libproxy.mit.edu/login?url=http://search.ebscohost.com/login.aspx?direct=true&db=bth&AN=7223171&site=ehost-live

Jardini, D. R. (2000). Out of the blue yonder: The transfer of systems thinking from the Pentagon to the great society, 1961-1965. *Systems, experts, and computers: the systems approach in management and engineering, World War II and after*, 311–57.

Johnson, L. B. (1966, June). *296 - Memorandum on the Use and Management of Computers by Federal Agencies.* President of the United States. Retrieved 2017-01-25, from http://www.presidency.ucsb.edu/ws/index.php?pid=27677

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., & Pohlmann, N. (2013). Online controlled experiments at large scale. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 1168–1176). ACM. Retrieved 2017-02-07, from http://dl.acm.org/citation.cfm?id=2488217

Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery, 18*(1), 140–181. Retrieved 2016-01-20, from http://link.springer.com/article/10.1007/s10618-008-0114-1

Kraut, R. E., Resnick, P., Kiesler, S., Burke, M., Chen, Y., Kittur, N., ... Riedl, J. (2012). *Building successful online communities: Evidence-based social design.* Mit Press.

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., ... others (2009). Life in the network: the coming age of computational social science. *Science (New York, NY), 323*(5915), 721. Retrieved 2016-04-24, from http://www.ncbi.nlm.nih.gov/pmc/articles/pmc2745217/

Lewin, K. (1944). The dynamics of group action. *Educational leadership, 1*(4), 195–200. Retrieved 2016-12-05, from http://www.ascd.com/ASCD/pdf/journals/ed_lead/el_194401_lewin.pdf

Lomas, N. (2017, March). Twitter nixed 635k+ terrorism accounts between mid-2015 and end of 2016. *TechCrunch.* Retrieved 2017-03-28, from https://techcrunch.com/2017/03/21/twitter-nixed-635k-terrorism-accounts-between-mid-2015-and-end-of-2016/

Marrow, A. J. (1969). *The Practical Theorist: The Life and Work of Kurt Lewin.* Basic Books.

Marrow, A. J., & French, J. R. (1945). Changing a stereotype in industry. *Journal of Social Issues, 1*(3), 33–37. Retrieved 2016-12-11, from http://onlinelibrary.wiley.com/doi/10.1111/j.1540-4560.1945.tb02691.x/full

McGee, A. M. (2007). *"Please Mr. Machine, Give this to a Human to Read": Electronic Data Processing, Systems Management, and Great Society Idealism in the Social Security Administration, 1965 - 1974* (M.A.). University of Virginia.

Merrill, J. (1964). *The pushcart war.* Harper & Row.

Mills, C. W. (1951). *White collar: The American middle classes* (Vol. 3). Oxford University Press on Demand.

Nadworny, M. J. (1955). *Scientific management and the unions, 1900-1932; a historical analysis.* Harvard University Press.

National Industrial Recovery Act of 1933. (1933, June).

National Labor Relations Act of 1935. (1935, July).

Oakley, A. (1998, October). Experimentation and social interventions: a forgotten but important history. *BMJ*, *317*(7167), 1239–1242. Retrieved 2016-12-17, from http://www.ncbi.nlm.nih.gov/pmc/articles/PMC1114168/

Oakley, A. (2000). Experiments in knowing: Gender and method in the social sciences. Retrieved 2016-12-17, from http://philpapers.org/rec/OAKEIK

Parks, J. (2017, May). *Scaling Airbnb's Experimentation Platform – Airbnb Engineering & Data Science – Medium.* Retrieved 2017-05-12, from https://medium.com/airbnb-engineering/https-medium-com-jonathan-parks-scaling-erf-23fd17c91166

Popper, K. (1947). *The open society and its enemies.* Routledge.

Prahalad, C. K., & Ramaswamy, V. (2004). Co-creation experiences: The next practice in value creation. *Journal of interactive marketing*, *18*(3), 5–14. Retrieved 2017-01-17, from http://onlinelibrary.wiley.com/doi/10.1002/dir.20015/abstract

Puckett, C. (2010). Administering Social Security: Challenges Yesterday and Today. *Soc. Sec. Bull.*, *70*, 27.

Rosenblat, A., & Stark, L. (2016, July). *Algorithmic Labor and Information Asymmetries: A Case Study of Uber's Drivers* (SSRN Scholarly Paper No. ID 2686227). Rochester, NY: Social Science Research Network. Retrieved 2017-04-11, from https://papers.ssrn.com/abstract=2686227

Rossi, P. H., & Wright, a. J. D. (1984). Evaluation Research: An Assessment. *Annual Review of Sociology*, *10*(1), 331–352. Retrieved 2017-01-07, from http://dx.doi.org/10.1146/annurev.so.10.080184.001555 doi: 10.1146/annurev.so.10.080184.001555

Scheiber, N. (2017, April). How Uber Uses Psychological Tricks to Push Its Drivers' Buttons. *The New York Times.* Retrieved 2017-04-11, from https://www.nytimes.com/interactive/2017/04/02/technology/uber-drivers-psychological-tricks.html

Seltzer, W. (2010). Free speech unmoored in copyright's safe harbor: Chilling effects of the DMCA on the first amendment. *Harv. JL & Tech., 24*, 171.

Shadish, W. R., Cook, T. D., & Leviton, L. C. (1991). Donald T. Campbell: Methodologist of the experimenting society. *Foundations of program evaluation*, 73–119.

Taylor, F. W. (1914). *The principles of scientific management.* Harper.

Terranova, T. (2000). Free labor: Producing culture for the digital economy. *Social text, 18*(2), 33–58. Retrieved 2017-01-17, from https://muse.jhu.edu/article/31873/summary

Thakor, M. N. (2016). *Algorithmic detectives against child trafficking: data, entrapment, and the new global policing network* (Doctoral dissertation, Massachusetts Institute of Technology). Retrieved 2017-03-28, from https://dspace.mit.edu/handle/1721.1/107039

Thaler, R. H., & Sunstein, C. R. (2003). Libertarian paternalism. *The American Economic Review, 93*(2), 175–179. Retrieved 2017-01-15, from http://www.jstor.org/stable/3132220

Thaler, R. H., Sunstein, C. R., & Balz, J. P. (2014). Choice architecture. *The behavioral foundations of public policy.* Retrieved 2017-01-17, from http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2536504

Vaas, L. (2017, April). *Uber 'showing drivers and riders different fare estimates', says lawsuit.* Retrieved 2017-04-11, from https://nakedsecurity.sophos.com/2017/04/10/uber-showing-drivers-and-riders-different-fare-estimates-says-lawsuit/

Valentine, R. (1916, January). The progressive relation between efficiency and consent. *Bulletin of Taylor Society, 2*(1).

Wark, L. (2016, November). Inside Alphabet's Jigsaw, the powerful tech incubator that could reshape geopolitics. *Quartz.* Retrieved 2017-03-28, from https://qz.com/846836/inside-google-jigsaw-the-powerful-tech-incubator-that-wants-to-reshape-geopolitics/

White, J. (2012). *Bandit algorithms for website optimization. "* O'Reilly Media, Inc.".

Williams, W. (1971). *Social Policy Research and Analysis: The Experience in the Federal Social Agencies.* American Elsevier Publishing Company. Retrieved 2017-01-07, from https://www.ncjrs.gov/App/Publications/abstract.aspx?ID=16261

Yates, J. (1993). *Control through communication: The rise of system in American management* (Vol. 6). JHU Press.

Chapter3

Governance by Volunteer Moderators Online

For over forty years, designers have created internet platforms to rely on volunteer moderators to govern community groups online. Because these moderators and their communities carry out their own policies and collect their own data within platforms, they possess the potential for community-led policy experiments. In this chapter, I describe the work of volunteer moderators, examining the ways they are accountable to their communities, to platforms, and to other moderators. To describe this work, I spent time as an ethnographer with moderators on the reddit platform, a social news site that had over 148,000 moderator roles in July 2015. I show how moderators negotiate their relationships and their accountability with all three stakeholders in their everyday moderation work.

When Stephen Hawking agreed to an online question-answer session on the social news platform reddit, he may not have expected the level of abuse and harassment the conversation would attract. While some people asked questions about time, the universe, and artificial intelligence, others sent Hawking scatological insults and shared obscene jokes about his private life.

Stephen Hawking's experience is a common one. In the U.S., 47% of internet users have personally experienced some kind of online harassment, from insults to sexual harassment and death threats. The social cost of this harassment extends well beyond risks of physical harm; 27% of Americans report self-censoring to avoid harassment (Lenhart, Ybarra, Zickuhr, & Price-Feeney, 2016). While many people who experience online harassment withdraw from social support, they are not alone; 46% of Americans also report taking some action to support or defend someone who experienced online harassment (Lenhart et al., 2016).

The personal attacks and offensive jokes directed at Stephen Hawking were removed by volunteer moderators of r/science, the community of over 15 million subscribers who share and discuss peer-reviewed academic research on reddit. Other readers had already "downvoted" the comments, and once moderators determined that they violated the community's policies against abusive comments, they were removed. Moderation in this community is provided by over a thousand volunteers who are all faculty, students, or graduates of universities around the world. In July 2016, moderators removed over 130,000 comments or posts, banned 460 people from the community, and made 558 changes to the community's policy and planning documents.

When someone threatens, disparages, or otherwise harasses another person online, volunteer moderators offer one of three three kinds of authority governing that behavior. *Government regulations* about hate speech, threats of violence, child pornography, and copyright sometimes lead to court cases and content removals (Citron, 2009; Marwick & Miller, 2014). The *operators of online platforms,* intermediaries such as Facebook and reddit, also develop internal policies about unacceptable behavior (Gillespie, 2010; Citron & Norton, 2011). When people who use these platforms report behavior they dislike, paid staff review the reports and make decisions to remove content, ban people from the service, or escalate an issue to law enforcement (Chen, 2014; Crawford & Gillespie, 2014; Matias et al., 2015; Buni & Chemaly, 2016). Yet even as platforms take some responsibility to observe and govern the digital behavior of

up to billions of people, they struggle to develop scalable policies that can be consistently applied across jurisdiction, language, context, and culture (?, ?). *Volunteer moderators* provide the most local form of governance online. Hundreds of thousands of groups across the the social web including the moderators of r/science create policies (Butler, Sproull, Kiesler, & Kraut, 2002), monitor activity in their communities (Geiger & Ribes, 2010), carry out interventions, and report serious cases to platforms and law enforcement.

Is volunteer moderation a form of community governance or are moderators a free, unaccountable labor pool who extend their own power and the power of platforms in society? When making sense of the work of moderation, scholars have tended to think primarily in one of three ways. Scholarship on digital labor describes moderation as unwaged labor for commercial interests or free labor in peer production communities like Wikipedia (Terranova, 2000; Postigo, 2003; Menking & Erickson, 2015). Legal theorists and computer scientists describe moderators as civic leaders of online communities who build their own public spheres (Kelty, 2005; ?, ?); much of this scholarship outlines general strategies to structure governance work for fair and functional communities at scale (Butler et al., 2002; Grimmelmann, 2015). A third conversation draws from the sociology of participation to consider the social structures of those who acquire and exercise moderation power, finding that common tendencies toward oligarchy on platforms like Wikia (Shaw & Hill, 2014) may be necessary for the survival of online communities (Zhu, Kraut, & Kittur, 2014).

Even as scholars debate the nature of moderation work, online communities routinely define what it means to be a moderator in everyday settings: they dispute over moderator decisions, recruit new moderators, participate in elections, investigate corruption, offer mentorship, and share peer support. In these conversations, especially at moments of tension and transition, moderators negotiate their policies and their power with the communities they govern, other moderators, and the operators of the online platform. Nor is the power they exercise merely toward communities; moderators also possess the power to change platform design and influence the organizations that control a platform. This everyday work of defining moderation has implications well beyond any individual community's interests. On reddit alone, volunteer moderators create and enact policy that governs roughly six percent of Americans, work that fundamentally shapes our digitally-mediated social and political lives (Barthel, Stocking, Holcomb, & Mitchell, 2016).

By asking who moderators are accountable to, we can begin to imagine their role in the present and future of platform governance. In this chapter, I describe the everyday ways that moderators negotiate their role with a platform, with their communities, and with each other. In ethnographic fieldwork and observations across hundreds of communities, I describe the kind of power that moderators hold with each stakeholder, as well as the kind of accountability that they face with each one. I conclude with a case study of the reddit blackout of July 2015, a moment when reddit moderators, often with the support of their communities, organized a strike to change how the the reddit company structured and supported moderator governance. While some observers saw the blackout as evidence of moderator oligarchy, I argue that this dramatic moment reveals the kinds of accountability that moderators face from their communities and the collective power of communities and their moderators to hold a platform to account and influence its decisions.

Moderation Work

Volunteer moderators have played a fundamental role in social life online for over 40 years, from librarians in 1970s Berkeley looking after local message-boards (Bruckman, 1998) to today's Facebook group administrators (Kushin & Kitchener, 2009), Wikipedia arbitrators (Menking & Erickson, 2015), and reddit moderators. Although not all work of fostering community is carried out by formal moderators, people in these formal positions are founders, maintainers, content producers, promoters, policymakers, and enforcers of policy across the social internet (Butler et al., 2002). On many platforms, moderators also manage autonomous and semi-autonomous moderation software that work alongside them (Geiger & Ribes, 2010).

By delegating policy and governance power to moderators, platform operators reduce labor costs and limit their regulatory liability for conduct on their service while also positioning themselves as champions of free expression and cultural generativity (Gillespie, 2010). This governance work invites public scrutiny, which draws platforms into debates about their responses to flagged material (Crawford & Gillespie, 2014). However, when platforms delegate policy-making to their users, that scrutiny is faced instead by moderators, whose labor nonetheless upholds a platform's economic model.

The evolution of moderation over the history of reddit followed this longer

40 year pattern. When reddit's creators founded it in 2005 to be "the front page of the Internet," they developed an infrastructure for sharing and promoting highly-voted posts a single, algorithmically-curated page. After these algorithms regularly promoted pornography and other complicated, possibly illegal material, the platform created an alternative algorithmic space for "Not Safe For Work"(NSFW) material, calling it a "subreddit" one month later (Huffman, 2006). Over the next two years, the company started dozens of new subreddits, mostly to separate conversations in different languages. In Jan 2008, after its acquisition by Condé Nast and 10 months after introducing advertising, the company launched "user-controlled subreddits." Before then, users could join official company subreddits, reporting spam and abuse directly to the company through a flagging system. Now they could create their own public and private subreddits, taking action themselves to "remove posts and ban users" (Huffman, 2007, 2008).

By June 2015, reddit was one of the largest social platforms online. That month reddit received over 160 million visitors,[1] roughly half of the number of active Twitter users in the same period.[2] To maintain social relations at that scale, the platform relied on nearly one hundred fifty thousand moderator roles[3] for over fifty-two thousand monthly active subreddits.

Just who do moderators work for? Your answer to this question may depend on whether you see moderators as laborers, community servants, or oligarchs.

Moderation as Free Labor in a Social Factory

Digital labor scholarship on the work of moderators foregrounds their relationship with online platforms, theorizing the role of moderators' volunteer work within platform business models. Among examples in open source and free culture, this scholarship also frequently refers to labor organizing by community leaders (essentially moderators) of AOL chatrooms and other communities in the 1990s. Initially eager to offer moderation work in exchange for

[1] `http://web.archive.org/web/20150703012219/http://www.reddit.com/about` Accessed 3 July 2015

[2] `http://web.archive.org/web/20150704143845/https://about.twitter.com/company` Accessed 4 July 2015

[3] Many accounts have multiple moderator positions, and some use "throwaway accounts" and "alts" on reddit (Leavitt, 2015). Consequently, this number over-estimates the number of people involved.

discounts, credit, and other perks, some of the 14,000 "community leads" came to see their work as unpaid labor. Moderators filed a class action lawsuit in 1999, prompting an inconclusive U.S. Department of Labor investigation. The community leaders eventually won $15 million dollars from AOL in a 2008 settlement (Postigo, 2009; Kirchner, 2011).

In her analysis of labor organizing by AOL moderators, Terranova points out that this freely given labor comprises an arrangement where people carry out self-directed cultural and social work that produces the value extracted by platforms. For Terranova, the "free labor" of platform production is something that is both "not financially rewarded [by platforms] and willingly given [by users]." (Terranova, 2000).

In a series of articles on the AOL lawsuit, Postigo explores the nature of the delicate symbiosis between platforms and moderators by observing the factors that led this arrangement to collapse. Postigo observes that the gift of volunteer time by AOL moderators was inspired by the "early Internet community spirit" found in "hacker history" and in "the academic, collaborative efforts that shaped the Internet" in the 1960s, 70s, and 80s. Yet some also took on the role to grow their technical skills or gain the discounts initially offered to volunteers. As AOL grew, the company began to formalize and control the relationship with their community leaders through communications, software, and compensation structures. No longer allowed the autonomy to imagine themselves as cultural gift-givers, the community leaders re-imagined themselves as mistreated employees and sued the company. Postigo describes their labor organizing as an effort to "stake out new occupational territory" for "community making" on the internet, an example of people who were "breaking out of the 'social factory'" that Terranova put forward (Postigo, 2003, 2009).

Terranova and Postigo rightly draw attention to the co-dependence of many online platforms with the substantial uncompensated labor that continues to support them. Community management is now more common as a paid position, but the majority of the labor continues to be unpaid. Theories of digital labor offer clarity on the challenges of creating a "profitable business," through volunteer labor, as Adrian Chen phrased it in the New York Times. Yet in many ways, the reddit blackout defies explanation by these theories. Moderators did not attempt to stake out their work as an occupation, nor did they demand compensation. Instead, they leveraged reddit's dependence on advertising to force the company to better meet their needs and those of their communities.

As Centivanny argues, the reddit blackout was a social movement focused on company policy, a moment where the dependence of a platform on volunteer labor was deployed to achieve aims with as many civic dimensions as economic ones (Centivany & Glushko, 2016).

Moderation as Civic Participation

The work of moderation online is the work of creating, maintaining, and defining "networked publics," imagined collective spaces that "allow people to gather for social, cultural, and civic purposes" (boyd, 2010). While social platforms offer technical infrastructures that constitute these publics, the work of creating and maintaining these imagined spaces is carried out in many everyday ways by platform participants and moderators.

Butler and colleagues call the work of moderation "community maintenance," drawing attention to the "communal challenge of developing and maintaining their existence." They compare these communities to neighborhood societies, churches, and social movements. Writing about the details of community work online, Butler and colleagues draw attention to the benefits of affiliation and social capital. Where Terranova and Postigo see labor in service of platform business models, Butler and his colleagues describe community maintenance as a service to the community itself (Butler et al., 2002). Consequently, their survey research imagines moderation similarly to any community work. Aside from the unique challenges of tending community software, the mailing list moderators studied by Butler support their communities by recruiting newcomers, managing social dynamics, and participating in the community.

As online harassment has grown in prominence, scholarship on the role of moderators has drawn attention to their work to protect people's capacities to participate in publics. These volunteers create and manage technical infrastructures such as "block bots" and moderation bots to filter "harassment, incivility, hate speech, trolling, and other related phenomena," argues Stuart Geiger. These volunteer efforts see moderation as "a civil rights issue of governance," where marginalized groups deploy community infrastructure to claim spaces for conversation, community, and support (Geiger, 2016).

While these civic perspectives on moderation acknowledge the role of platforms, they foreground the relationship between moderators and the publics they are responsible for. The labor of moderators does sustain platform economies,

yet the work itself is most obviously concerned with the specific communities they govern.

Moderation as Oligarchy

Even as a moderator's work supports their community, the power of individual moderators is defined and managed by other moderators who gate-keep the process of taking on the role. A third perspective on moderation work examines ways that moderation work is socially structured and how moderator interests can diverge from the goals of their communities.

Early theories of leadership development in online communities imagined a continuous "reader to leader" process where more active participants gain greater responsibility over time in a regular churn (Preece & Shneiderman, 2009). However, longitudinal research by Shaw and Hill has shown online communities to be much more like other voluntary organizations, where "groups of early members consolidate and exercise a monopoly of power within the organization as their interests diverge from the collective's." Across 683 Wikia wikis, they find support for this "iron law of oligarchy," showing that on average, a small group does come to control the positions of formal authority as a wiki grows (Shaw & Hill, 2014). Yet where Shaw and Hill see oligarchy, others see experience necessary for online communities to flourish. Also studying Wikia, Zhu and colleagues interpreted similar findings to argue that communities whose leaders also lead other communities are more likely to survive and grow (Zhu et al., 2014). In all these cases, experienced and powerful moderators control the process for others to gain and maintain their positions. Anyone seeking the role must negotiate that position with other moderators as well as their community and the platform.

These highly-compatible studies vary widely in their framing of moderation because their findings are purely quantitative. Stable leadership might support oppression if moderators routinely ignore the interests and demands of their communities. On the other hand, long-lasting communities with stable groups of moderators might also succeed through processes that invite communities to exercise substantial power in community affairs. Because these quantitative studies limit their observations to narrow variables of moderator tenure, moderator experience, and community longevity, they do not offer any evidence on the relationships between moderators and their communities.

Standpoint and Methods

I came to this research after leading a team to study efforts by Women, Action, and the Media (WAM!), an NGO offering support to people experiencing harassment on Twitter (Matias et al., 2015). The volunteers who reviewed harassment reports and advocated the cases to Twitter were criticized from multiple directions. Some argued that these advocates represented a step backward for progress on online harassment, taking on work that Twitter should be paying for (Meyer, 2014). Others called it a dangerous form of censorship (Sullivan, 2014). As our team studied the work of reviewing and responding to Twitter harassment, I was deeply moved by the overwhelming amounts of labor and personal risk taken by the harassment reviewers. Volunteers handled cases at all hours and became harassment targets themselves. One volunteer dropped out after experiencing severe post-traumatic stress. Furthermore, WAM! also needed to manage their relationship with Twitter to retain the privilege of supporting harassment receivers and maintain a public voice on the company's policies.

My fieldwork with reddit moderators began at a time when I was trying understand the many-sided scrunity that WAM!'s harassment reviewers had faced. Volunteer responders might be unpaid, but they were a privately selected group with substantial power over others. Their work served platform operators who could remove them at will. They also served and governed users, who pressured them to share and justify their actions. As I spent time with reddit moderators, I watched them respond to similar questions from these multiple sides, a position many moderators had been negotiating for years.

To study the accountability of moderators conduct to platforms, communities, and each other, I carried out participant observation, content analysis, interviews, and trace data collection on the social news site reddit over a four-month period from June through September 2015, with followup data collection through February 2016. Collected content includes 10 years of public statements by the company, 90 published interviews by moderators of other moderators, statements by over 200 subreddits that joined the blackout, over 150 subreddit discussions after concluding participation in the blackout, and over 100 discussions in subreddits that declined to join the blackout.[4] I also conducted trace analysis of moderator roles in the population of 52,735 active

[4]Quotations from subreddit discussions have been obfuscated to protect participant privacy

subreddits. Finally, I held semi-structured interviews with 14 moderators of subreddits of all sizes, including those on both sides of the blackout. Interviewees included moderators of "NSFW" subreddits only available to users 18 years or older, as well as more widely accessible subreddits. Moderators of subreddits allegedly associated with hate speech declined to participate.

In this chapter, I focus on moments of tension and transition that brought debates over the meaning of moderation to the fore, including disputes over moderator decisions, the process of becoming a moderator, transitions of leadership, conflicts between communities, crises of legitimacy, the work of starting new communities, debates over compensation, and collective action during the reddit blackout of July 2015.

Disputing and Justifying Moderation Decisions with Communities

When someone's contribution to reddit is removed by moderators, it can often come as a surprise. Since many participants engage primarily with the platform's aggregated feed, they may not be aware that the posts they submit are subject to a subreddit's community policies (Massanari, 2015). Responses to moderation decisions are often received through "modmail," a shared inbox for each subreddit's moderators. Complaints often include moderation policy debates, profanity, racist slurs, and threats of violence.

Even when moderators ignore the complaints, these disputes shape the language the moderators use to describe their roles as dictators, martyrs, janitors, hosts, taste-makers, and policymakers.

Some moderators describe themselves as "dictators," arguing that the power they exercised needed no justification. In these communities, "the top mod makes all the decisions, usually because s/he created the sub." Those who complain are urged either to accept moderator power or stay away.

Moderators of subreddits dedicated to marginalized communities sometimes explain themselves as defenders. One moderator described the former moderator of a gender minority subreddit as a "martyr, angry and whirling and ready to give hell to anyone who dared to cross her or to threaten her communities." When adopting the figure of a defender, moderators draw attention to the moral and political justifications for their exercise of power.

Other moderators adopt language from hospitality or service labor, de-

scribing themselves as "hosts" and "janitors." These analogies de-politicize their role. Describing themself in this way, one moderator argued that "my subreddits belong to my communities, I just happen to help out by cleaning up." Reflecting on the accusations and complaints they receive, another moderator explained:

> It seems like it's some sort of important position, while it's actually just janitoral work...the degree of accusations, insults, abuse and unreasonable complaints from the politically interested is extreme...it's janitorial when you remove hundreds of comments that just say "kill yourself blackie"

When I asked moderators whether the language of janitor also implied a labor critique towards the reddit company, they disagreed. One described the language of janitor as "a response to complaints about conspiracies, censorship, etc" rather their relationship to the company.

Many moderators describe themselves as taste-makers when explaining their decisions about what to remove. In one subreddit dedicated to shocking material, moderators expressed disappointment over the lack of nuance and quality in submitters' sense of the truly shocking. For example, one moderator claimed that too many submitters are shocked by images of nudity, violent injury, or death; moderators considered these too commonplace for inclusion. These moderators described themselves as taste-makers for their communities: "we are fucked up, but in a courtesy sniff kinda way that you're ok with sharing with your friends."

Some moderators respond to complaints of censorship by drawing inspiration from the language of governance. These subreddits describe their decisions in terms of "policies" and sometimes produce transparency reports of moderation actions. One subreddit described its transparency report as a response to participant complaints, an effort "towards improving user-moderator relations." [5] Their five page report offered an empirical response to common complaints received by moderators of this 10 million subscriber community. Several other large subreddits publish aggregated transparency reports, with some sharing public logs of every action taken by the group's moderators. By

[5] https :// www .reddit .com / r / science / comments / 43g15s / first _transparency_report_for_rscience/

publishing transparency reports, moderators position themselves as civic actors accountable to their communities. The reports deflect criticism while also inviting evidence-based discussions of moderation practices.

The language of governance is also used by reddit participants who investigate and analyze moderator behavior. One interviewee described investigating and "exposing" a moderator for encouraging reddit users to share sexual photographs of minors. The investigators organized a press campaign to pressure the company, who then shut down the subreddit involved (Morris, 2011). In another case, participants accused a large technology subreddit's moderators of censoring political discussions. To support these accusations, one reddit user conducted data analysis of the subreddit's history, creating charts that showed a sharp cutoff in discussions of surveillance and other political topics. The moderators' accusers argued that the subreddit lacked "accountability" and "transparency." After the reddit platform sanctioned the subreddit amidst substantial international press coverage, the moderators also invoked the language of governance, making a formal public statement that "the mods directly responsible for this system are no longer a part of the team and the new team is committed to maintaining a transparent style of moderation." ("Reddit downgrades technology community after censorship", 2014; Collier, 2014)

Moderator Internships, Applications, and Elections

The practical work of recruiting and choosing new moderators requires people to define what it means to be a moderator. Since a subreddit's current moderators control the reddit software's process of appointing new moderators, would-be moderators must justify themselves and their ideas of the work to their would-be peers. Likewise, current moderators invest substantial labor into the work of admitting new moderators. At these moments of transition, democratic, oligarchic, and professional notions of moderator work come into tension as subreddits negotiate who should select the leaders and what qualities they should demonstrate.

Among those interviewed, moderators gained their positions through wide range of means. One was added by a school friend who needed extra help. Others were invited to be moderators after demonstrating substantial participation in the subreddit's affairs. One was made a moderator in appreciation of their role to expose the scandal over sexual images of minors. Some were recruited

for their expertise at operating the reddit platform software. Yet many subreddits also operate formal structures for adding moderators, systems that draw from the language of the workplace and the public sector.

Many subreddits hold a formal application process for becoming a moderator. In the simplest versions, interested parties fill out an interview form, noting their timezone and availability, describing their moderation experience, listing their skills, and explaining their reasons for applying. One popular subreddit received 600 applications in one recruitment effort, identified a shortlist of 60 applicants to interview, and chose from the shortlist. The process from call to selection can take from weeks to over a month.

While moderator teams sometimes take final responsibility for selecting new moderators—what Shaw and Hill call oligarchy—some subreddits open the final selection to subscribers. The reddit platform doesn't support ballots, so subreddits have developed their own voting systems. Speaking about the elections in one subreddit for a minority group, a moderator explained, "I got one ballot, just like every one else." Yet especially with elections, moderators still felt responsible to to filter possible nominees lest the wrong person become elected. The same moderator explained that public opinion wasn't appropriate for nominating candidates since it risked reinforcing prejudice: "lots of people who can't be bigots so much anymore [due to social pressure] have found that they can still target [minority group] and nobody seems to mind."

If voting software supplies infrastructure for democratic notions of moderation, the moderator job board on reddit offers infrastructure for more oligarchic forms. This subreddit publishes moderation opportunities alongside "offers to mod." Postings routinely offer arguments on the nature of moderation work, such as the disinterested approach to moderation offered in one job listing for a community with frequent conflicts:

> I'm looking for an impartial moderator, who doesn't belong to [organization], and who doesn't hold a specific view on it. Must have:
>
> - been on reddit for at least 2 years
> - moderating experience
>
> The sub is an open platform to discuss [topic], but prejudiced comments aren't allowed.

Soon after the primary moderator posted this message, community members, who had noticed the listing, added objections: "Seriously? We have posted so many requests for mods to that sub. We have even posted solutions that result in a very balanced 3 party system." This three-party system would entail asking participants from each faction to choose a representative who would take a role as moderator of the subreddit. The participants who proposed the three-party system accused the poster of delinquency and argued strongly against the idea of disinterested, objective moderation: "Anyone without knowledge on the subject will be unable to effectively moderate the sub." After an extended discussion, the moderator accepted their proposal, and the "three party system" was still in place over one year later.

Even democratic subreddits emphasize previous experience when selecting moderators, leading many to seek and tout their moderation "résumé." Since a medium-to-large subreddit is unlikely to accept applicants with limited experience, some subreddits grow their labor pool by offering "internships" and other entry-level moderation opportunities. /r/SubredditOfTheDay, which publishes original content every day, offers a two-month internship for people seeking moderation opportunities. Interns agree to write 6 original posts that feature interviews with the moderation teams of other subreddits. Those who finish the internship period are made full moderators, and they also gain opportunities to moderate other subreddits.

Among large subreddits that admit inexperienced moderators, newcomers are sometimes admitted in cohorts and offered mentorship that can last for several months. As new moderators demonstrate their capabilities, they are given greater moderation powers upon election or appointment. Several large subreddits operate internal promotion structures that formalize responsibilities at each rank and offer documented criteria for career advancement in moderation.

Crises in Legitimacy and The Removal of Moderators

In technical terms, only two parties can remove a moderator from their position on reddit. Platform employees, known as "admins," occasionally remove moderators if they are convinced that the moderator was inactive or abusing their power. Moderators with greater seniority also possess the power to remove those within the same community who were appointed more recently.

In an interview, one moderator described a "coup attempt" by moderators who systematically removed others who disagreed with their political views. Another moderator noticed the attempt in time and reinstated the ejected moderators. In another case, the sibling of someone who moderated a 30,000 subscriber group compromised their reddit account, took charge of the subreddit, and only restored it upon receiving threats of violence. Many moderators, especially those of large or contentious subreddits, pay close attention to their personal information security to protect against such takeovers. Platform employees will also occasionally take action to restore a subreddit's moderators when asked.

Moderators are more commonly removed for failing to perform their role. In some cases, would-be moderators appeal to the platform, who offer a process for requesting moderation of "inactive" subreddits. In other cases, a moderator loses their legitimacy to govern. In these cases, community participants sometimes pursue the person they mistrust, incessantly mocking their pronouncements and questioning their decisions. Such cases tend to conclude with a post from the moderator announcing their resignation, or a post from other moderators announcing that the offending moderator has been removed.

Moderator Compensation and Corruption

In 2012, a moderator of three of the largest subreddits posted links to an online news outlet after he had been hired as a social media advisor by the publisher's marketing firm (Morris, 2012). In response, the reddit platform banned the user and added a rule against third party compensation. Moderators also receive substantial scrutiny and criticism from their communities for alleged "corruption."

In one case, someone sent messages on the reddit platform to "a few dozen" moderators, offering compensation for help promoting their content. When some moderators reported the offer to reddit, employees investigated the private messages of everyone who received the offer. When the employees noticed that some moderators had responded positively, the company banned their accounts, including moderators of some of the platform's largest, most popular NSFW subreddits (Martinez, 2013). In 2015, a large gaming company asked moderators to remove links to material that could not legally be published, offering moderators early access to an upcoming Star Wars game in

exchange for their help. When one moderator reported the relationship to reddit employees, the others removed the moderator for a time, until they themselves were banned by reddit for accepting a "bribe." A reddit representative explained that the gaming company should have used alternative channels to address illegally-shared material (Khan, 2015). In another case, a mobile phone manufacturer offered "perks" to moderators of a subreddit that commonly discussed their products. In exchange, the company asked that its employees be made moderators. To protect themselves from community disapproval or platform intervention, moderators reported the request to reddit and posted the offending messages for discussion by their community (Farrell, 2015).

In interviews, moderators were insistent that they did not seek compensation, arguing that news articles that focused on their unpaid status failed to understand the nature of their work. One interviewee brought up the AOL community leader program, arguing that reddit moderators were different because they weren't managed as closely as the AOL volunteers. This independence was important to many moderators, including one who claimed, "I don't think I work for reddit. I run communities and reddit is the tool I use to do that." Yet at the time of the reddit blackout, moderators also felt ignored by the company behind these 'tools.' One explained that "it doesn't help when the site you are on doesn't appreciate/recognize/care about the cumulative thousands and thousands of hours the mods put in to make their site usable."

Starting Subreddits and Governing Moderator Networks

While some new subreddits are created to support a community that migrated to reddit from elsewhere, many moderators describe "founding" a subreddit and developing a growing community over time. Yet even the work of creating new subreddits requires managing the expectations of platform operators, moderators, and community participants. In interviews, I observed these negotiations among relationship-themed subreddits and networks of subreddits.

> I never intended to moderate a NSFW subreddit. It blew me away the community want for it

Relationship subreddits offer listings of people who are looking for conversations, pen-pals, and relationships, sometimes sexual, but often not. When one moderator started a group for users of a mobile messaging system, their

goal was to help newcomers on the messaging platform "find more people to chat with," whatever age. As the subreddit grew, participants continued to post requests for relationships and conversations that could be illegal for minors. These "dirty" relationship requests also put the subreddit at risk of intervention from reddit employees. Rather than designate the subreddit "NSFW," which would limit minors from accessing the group, the moderator created a parallel subreddit for "dirty" relationship matching. By splitting the conversation, the moderator found a way to meet community expectations while also protecting the primary subreddit from platform intervention.

Creators of new subreddits also work to comply with the expectations of other moderators, especially if they seek to join a subreddit "network." These networks are jointly-managed collections of subreddits that share moderators and a common governance structure. Some networks specialize in a particular kind of content. Several offer inspiring general-interest photography; others share celebrity pornography. Some networks adopt a structure akin to city states. To join the network, a moderator must grow their subreddit to a minimum size, institute a set of network-designated policies, and convince a "champion" within the network to advocate for their inclusion. These champions also help new network members comply with the network's requirements. New subreddits are inducted by vote from the moderators. At the time of writing, the largest two networks included 169 and 117 constituent subreddits, although networks also occur at smaller scales.

One network stopped accepting new subreddits after participants in a newly-added subreddit began "doxing" reddit users—a practice of publishing the addresses and phone numbers of people they disliked:

> one time we added a sub, vetted them, once we approved them, they started posting information on reddit users, so it looked like [the network] had approved doxxing, which was one of the two things that could get us banned [by the company].

Rather than risk reprisals from the platform operator, the network dissociated itself from the offending subreddit and halted all new applications. To address future risks, they required all groups to accept a lead moderator from the network's central leadership, to keep "everyone pointed in the same direction."

Figure 3-1: "Life of a Mod" comic by former moderator Daniel Allen, /u/solidwhetstone

Acknowledging Moderators' Position With Platform, Community, and Other Moderators

Two regularly shared comic strips by former moderator Daniel Allen remark directly on the work that moderators must do to manage their relationships with their communities, other moderators, and the reddit platform. The first 'life of a mod' comic strip presents moderators as people who carry out a wide range of community care for little appreciation. In the comic, moderators are janitors, referees, police, educators, and artists (Figure 3-1). The second presents the "Life of a Secret Cabal Mod," drawing attention to the accusations of oligarchy that moderators receive. The heading of each panel includes a common accusation towards moderators. The illustration beneath each heading offers an alternative explanation for the behavior that attracts accusation. For example, when one moderator helps another learn to remove what they see as hate speech, they could be accused of conspiring to silence dissent. When platform employees share software updates and moderators pass on commu-

Figure 3-2: Details from "Life of a Secret Cabal Mod" comic by former moderator Daniel Allen, /u/solidwhetstone

nity complaints to the company, they might also be accused of collusion (Figure 3-2). By drawing attention to the complicated negotiations that moderators conduct in multiple directions, Allen's comics themselves make a case for how those parties should see moderators.

Accountability and Influence in the reddit Blackout

On 2 July 2015, volunteer moderators of over two thousand two hundred "subreddit" communities on the social news platform reddit effectively went on strike. Moderators disabled their subreddits, preventing millions of subscribers from accessing basic parts of the reddit website. The "reddit blackout," as it became known, choked the company from advertising revenue and forced reddit to negotiate over moderators' digital working conditions. The company, already struggling with pressure from racist and regularly-harassing groups that it had recently banned, conceded to moderator demands within hours. Management allocated resources to moderator needs, CEO Ellen Pao resigned one week later, and within two months, the company had hired its first Chief Technical Officer, partly to improve the platform's moderation software (Olanoff, 2015).

Even as the blackout surfaced anxieties about the responsibilities of digi-

tal platforms to their volunteer workers, it also led many to question the legitimacy of moderators' governance role. Some moderators were censured or even ejected by their subreddits for joining the blackout without consulting their communities. Conversely, many moderators were pressured to join the blackout through subreddit-wide votes and waves of private messages.

Three weeks later, in a New York Times Magazine article on the word "moderator," Adrian Chen wrote:

> The moderator class has become so detached from its mediating role at Reddit that it no longer functions as a means of creating a harmonious community, let alone a profitable business. It has become an end in itself—a sort of moderatocracy (Chen, 2015)

Chen accurately recognized the stakeholders with whom moderators negotiate their role. While Chen describes the blackout as an abdication of moderator responsibilities, their experience of the blackout was much the opposite. While moderators hold some power with the platform, reddit participants, and other moderators, they also serve each of those masters. The interplay of these stakeholders becomes apparent when attempting to make sense of the reddit blackout, which was partly a labor dispute, partly a collective action from communities demanding better moderation, and also a coordinated effort by a group of organized moderators to gain expanded abilities serve their communities. By examining moderators' blackout decisions and community reactions after the blackout, I show how moderators managed those negotiations.

Deciding to Join the Blackout

The reddit blackout was precipitated when the company dismissed an employee who had consistently offered direct support to moderators in some of the site's most popular discussions: live question-answer sessions of the kind that Stephen Hawking joined, called Ask-Me-Anything threads (Isaac, 2015). Moderators of the /r/IamA subreddit described being caught off guard when the employee was dismissed in the middle of a live Q&A. When the community disabled their subreddit to decide their response (Lynch & Swearingen, 2015), other moderators of large subreddits took note. To these moderators, the company's failure to coordinate the transition with moderators was another sign of its neglect of moderator and community needs. Moderators had already been attempting

to convince the company to improve moderator software and increase its coordination with communities. In interviews, moderators explained that moderators of the largest groups had previously dismissed the idea of blacking out to make their needs clearer to the company. But "after she was fired, the idea came up again, [and] no one was really against it." These moderators described the blackout as a tactic that might give greater leverage to company employees who routinely advocated for moderator interests. When other moderators observed the behavior of these large groups, many joined the blackout, leaving messages on their subreddits expressing "solidarity" for moderators affected by the blackout.

Even as moderators discussed the blackout with each other, they also negotiated pressures from their communities over the decision to join the blackout. In interviews, moderators described receiving large volumes of private messages from participants that urged them toward or against the blackout. In response, many posted discussion threads asking for community opinions or announcing their decisions. In one post, a moderator apologized for "the inconvenience of going dark" and explained:

> I did get messages from people. The more I watched and saw more and more subs going down, I figured it was worth sending a message [to the platform]. We had kind of a mod vote and decided to black out.

Community interests were considered in many moderator decisions. One group of gaming-related subreddits, whose moderators see it as an "island just barely within reddit" concluded that joining the blackout would "punish our users who don't know or don't care about reddits politics." Yet they still faced pressure from many their community to join the blackout: "we eventually released the statement after we received dozens of modmails and posts on both subreddits."

Some moderators invited their communities to vote on participation in the blackout. In many cases, moderators followed the results of community votes. Yet networks of moderators did not always agree with their communities. In one subreddit in a subreddit network, one moderator held a vote that came out in favor of the blackout. The rest of the network stayed active; moderators more central to the network described the vote as a "rogue faction" and ignored it. Instead, they issued a proclamation that the entire network would

stay out of the protest. Elsewhere, one moderator described their community vote as a way to distract those who were clamoring for the blackout, gaining time for moderators to reach a collective decision. Many moderators and participants questioned the legitimacy of the votes that did occur, guessing that the results might be skewed by influxes of reddit users beyond their community who wanted to influence a community's decision.

Across these situations, moderators faced the same three questions: what would their actions say to the platform, to other moderators, and to their communities? The effect of the blackout on moderators' work would not be constrained to their relationship with the company—it would affect every other relationship in their everyday moderation work.

Defending Decisions After the Blackout

Moderators faced the consequences of their decisions once the blackout concluded. When the platform operators quickly ceded to moderator demands, many declared victory. Community and moderator reactions were more complex. While some subreddits systematically removed any mention of the blackout, it was more common for moderators to post a discussion explaining what had happened.

Especially for subreddits that were disabled for the entire weekend, this conversation could be heated. Only a small number of participants might notice a vote called at the moment of decision; many more would feel the effects of a blacked-out community. At these moments, moderators often defended themselves by referring to these votes. "You're all upset about the blackout decision. Which is silly. If you were upset why didn't you raise your concerns?" one wrote. In other cases, moderators assigned responsibility to a single moderator acting alone. Sometimes, they removed that person from the moderation team or encouraged them to resign.

In many of these discussions, moderators expressed support for the blackout, explained the reasons one might join the protest, and also apologized to their communities. These statements positioned moderators as supporters of the blackout while also defending themselves from community critiques. One recipe-sharing subreddit moderator took a compromise position by briefly joining the blackout and then re-opening in advance of July 4th U.S. Independence Day parties. They expressed their "full support" for the other moderators, drew

attention to an overwhelming community vote to black out, and then wrote an apology: "we are deeply sorry for the outage. Things need to change on reddit, and this was our best way to let them know our demands." These conversations reveal the very real accountability that moderators face from their communities, even in the few situations where moderators acted alone.

Governance by Volunteer Moderators in an Open, Experimenting Society

Volunteer moderators like those on reddit hold what may be the most accountable position of any form of platform governance. Legal responses to online harassment are rare (Citron & Norton, 2011) and platforms hide their policy work behind software interfaces (Buni & Chemaly, 2016; Crawford & Gillespie, 2014), but volunteer moderators hold visible user accounts. Communities can pressure them, platforms can remove them, and other moderators structure and influence their work. Scholarly perspectives on moderation are right to draw attention to these different stakeholders, but a clearer account of moderation work should attend to all three at once, just as moderators must always do. All three forces acculturate a moderator to their ever-changing position, from the application process to the moment they step down or are removed. From the most common dispute over a single comment removal to collective actions that make international news, the meaning of moderation is defined and defended with all three stakeholders.

In the first chapter of this dissertation, I described Donald Campbell's vision of an open, experimenting society where the people who are governed have routine power to influence policy experiments, dispute the findings, and influence policy decisions. I also argued that in our current era, where platforms are governing our social lives and large numbers of randomized trials are routine, we have an opportunity to avoid the authoritarian potential of platforms by developing an experimenting society for platform governance. In this hypothetical society, where participation in experiments forms a basic part of civic life, disputatious communities would imagine their own policy ideas, test them with community input, publicly criticize every aspect of the results, and develop compromises partly informed by evidence they developed themselves.

Any effort to implement the idea of an experimenting society could take one of three paths. A utopian endeavor might try to establish a new set of

communities founded on the principles of an experimenting society. A political project would attempt to convince platform leaders and researchers to renegotiate their relationship with the public. A socio-technical effort would seek out communities that already possess some characteristics of an open society and introduce experimentation as another capability in their governance repertoire.

Compared to other forms of governance power on platforms, the work of volunteer moderators is already often subject to the characteristics of an open society: criticism, compromise, and the ability to remove policies or leaders through public pressure. Moderator accountability has been a recurring pattern in a 40 year history of volunteers being invited, elected, and chosen into governance positions online. Nor is volunteer moderation unique to for-profit platform arrangements; moderators of non-profit platforms such as Wikipedia face a similar set of stakeholders to maintain their roles. In the rest of this dissertation, I build on what I learned about community governance by supporting moderators and their communities on reddit to conduct their own policy experiments.

References

Barthel, M., Stocking, G., Holcomb, J., & Mitchell, A. (2016, February). *Seven-in-Ten Reddit Users Get News on the Site* (Tech. Rep.). Pew Research Center for Journalism & Media. Retrieved 2017-04-02, from `http://www.journalism.org/2016/02/25/seven-in-ten-reddit-users-get-news-on-the-site/`

boyd, d. (2010). Social Network Sites as Netowrked Publics: Affordances, Dynamics, and Implications. In *Networked Self: Identity, Community, and Culture on Social Network Sites* (pp. 39–58). Routledge.

Bruckman, A. (1998). Finding one's own in cyberspace. *High Wired: On the Design, Use, and Theory of Educational MOOs. Ed. Cynthia Haynes and Jan Rune Holmevik. Ann Arbor, MI: U of Michigan P*, 15–24. Retrieved 2015-09-23, from `http://cumincad.architexturez.net/system/files/pdf/59c3.content.pdf`

Buni, C., & Chemaly, S. (2016, April). The secret rules of the internet. *The Verge*. Retrieved 2017-02-05, from http://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech

Butler, B., Sproull, L., Kiesler, S., & Kraut, R. (2002). Community effort in online groups: Who does the work and why. *Leadership at a distance: Research in technologically supported work*, 171–194.

Centivany, A., & Glushko, B. (2016). "Popcorn Tastes Good": Participatory Policymaking and Reddit's. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 1126–1137). New York, NY, USA: ACM. Retrieved 2016-08-29, from http://doi.acm.org/10.1145/2858036.2858516 doi: 10.1145/2858036.2858516

Chen, A. (2014, October). The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed. *WIRED*. Retrieved 2017-02-07, from https://www.wired.com/2014/10/content-moderation/

Chen, A. (2015, July). When the Internet's 'Moderators' Are Anything But. *The New York Times Magazine*. Retrieved 2016-02-03, from http://www.nytimes.com/2015/07/26/magazine/when-the-internets-moderators-are-anything-but.html?_r=0

Citron, D. K. (2009). Law's expressive value in combating cyber gender harassment. *Michigan Law Review*, 373–415. Retrieved 2015-06-26, from http://www.jstor.org/stable/40379876

Citron, D. K., & Norton, H. L. (2011). Intermediaries and hate speech: Fostering digital citizenship for our information age. *Boston University Law Review*, *91*, 1435. Retrieved 2015-06-24, from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1764004

Collier, K. (2014, April). Reddit's r/technology has a secret list of about 50 words you can't use in headlines. *The Daily Dot*. Retrieved 2016-09-01, from http://www.dailydot.com/news/reddit-technology-banned-words/

Crawford, K., & Gillespie, T. L. (2014, August). What is a Flag for? Social Media Reporting Tools and the Vocabulary of Complaint. *New Media & Society*. Retrieved 2014-10-20, from http://papers.ssrn.com/abstract=2476464

Farrell, N. (2015, September). *HTC Tried to Bribe a Reddit Moderator and got Burned... Hard.* Retrieved 2016-08-30, from http://www.cybernole.net/news/htc-tried-to-bribe-a-reddit-moderator-and-got-burned-hard/

Geiger, R. S. (2016, June). Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society, 19*(6), 787–803. Retrieved 2016-08-29, from http://dx.doi.org/10.1080/1369118X.2016.1153700 doi: 10.1080/1369118X.2016.1153700

Geiger, R. S., & Ribes, D. (2010). The work of sustaining order in wikipedia: the banning of a vandal. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work* (pp. 117–126). ACM. Retrieved 2016-08-28, from http://dl.acm.org/citation.cfm?id=1718941

Gillespie, T. (2010). The politics of 'platforms'. *New Media & Society, 12*(3), 347–364. Retrieved 2015-12-08, from http://nms.sagepub.com/content/12/3/347.short

Grimmelmann, J. (2015, April). *The Virtues of Moderation* (SSRN Scholarly Paper No. ID 2588493). Rochester, NY: Social Science Research Network. Retrieved 2015-06-24, from http://papers.ssrn.com/abstract=2588493

Huffman, S. (2006, January). *what's new on reddit: for those of you with a private office...* Retrieved 2015-09-25, from http://www.redditblog.com/2006/01/for-those-of-you-with-private-office.html

Huffman, S. (2007, March). *what's new on reddit: brace yourself, ads are coming.* Retrieved 2015-09-25, from http://www.redditblog.com/2007/03/brace-yourself-ads-are-coming.html

Huffman, S. (2008, January). *what's new on reddit: new features.* Retrieved 2015-09-25, from http://www.redditblog.com/2008/01/new-features.html

Isaac, M. (2015, July). Reddit Moderators Shut Down Parts of Site Over Employee's Dismissal. *The New York Times.* Retrieved 2015-09-23, from http://www.nytimes.com/2015/07/04/technology/reddit-moderators-shut-down-parts-of-site-over-executives-dismissal.html

Kelty, C. (2005, May). Geeks, Social Imaginaries, and Recursive Publics. *Cultural Anthropology, 20*(2), 185–214. Retrieved 2016-03-14, from http://onlinelibrary.wiley.com/doi/10.1525/can.2005.20.2.185/abstract doi: 10.1525/can.2005.20.2.185

Khan, Z. (2015, November). *EA Reportedly Bribed Star Wars Battlefront Reddit Mods.* Retrieved 2016-08-30, from http://www.playstationlifestyle.net/2015/11/12/star-wars-battlefront-reddit-mods-bribed/#/slide/1

Kirchner, L. (2011, February). AOL Settled with Unpaid "Volunteers" for $15 Million. *Columbia Journalism Review.* Retrieved 2015-09-23, from http://www.cjr.org/the_news_frontier/aol_settled_with_unpaid_volunt.php

Kushin, M. J., & Kitchener, K. (2009, October). Getting political on social network sites: Exploring online political discourse on Facebook. *First Monday, 14*(11). Retrieved 2016-02-04, from http://journals.uic.edu/ojs/index.php/fm/article/view/2645 doi: 10.5210/fm.v14i11.2645

Leavitt, A. (2015). This is a Throwaway Account: Temporary Technical Identities and Perceptions of Anonymity in a Massive Online Community. In *Proceedings of the 18th ACM Conference on Computer Supported Cooperative Work & Social Computing* (pp. 317–327). ACM. Retrieved 2015-09-25, from http://dl.acm.org/citation.cfm?id=2675175

Lenhart, A., Ybarra, M., Zickuhr, K., & Price-Feeney, M. (2016, November). *Online Harassment, Digital Abuse, and Cyberstalking in America* (Tech. Rep.). New York, NY: Data & Society Research Institute.

Lynch, B., & Swearingen, C. (2015, July). Why We Shut Down Reddit's 'Ask Me Anything' Forum. *The New York Times.* Retrieved 2015-09-25, from http://www.nytimes.com/2015/07/08/opinion/why-we-shut-down-reddits-ask-me-anything-forum.html

Martinez, F. (2013, January). Top Reddit porn moderators banned for alleged bribes. *The Daily Dot.* Retrieved 2016-08-30, from http://www.dailydot.com/news/reddit-ban-porn-mods-nsfw-bribes/

Marwick, A. E., & Miller, R. W. (2014, June). *Online Harassment, Defamation, and Hateful Speech: A Primer of the Legal Landscape* (SSRN Scholarly Paper No. ID 2447904). Rochester, NY: Social Science Research Network. Retrieved 2014-10-02, from http://papers.ssrn.com/abstract=2447904

Massanari, A. (2015). # Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society,* 1461444815608807. Retrieved 2015-12-03, from http://nms.sagepub.com/content/early/2015/10/07/1461444815608807.abstract

Matias, J. N., Johnson, A., Boesel, W. E., Keegan, B., Friedman, J., & DeTar, C. (2015, May). Reporting, Reviewing, and Responding to Harassment on Twitter. *arXiv:1505.03359 [cs].* Retrieved 2015-09-23, from http://arxiv.org/abs/1505.03359 (arXiv: 1505.03359)

Menking, A., & Erickson, I. (2015). The Heart Work of Wikipedia: Gendered, Emotional Labor in the World's Largest Online Encyclopedia. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 207–210). ACM. Retrieved 2016-02-03, from http://dl.acm.org/citation.cfm?id=2702514

Meyer, R. (2014, November). The Good (and the Bad) of Twitter's New Bid to Stop Harassment. *The Atlantic.* Retrieved 2016-02-04, from http://www.theatlantic.com/technology/archive/2014/11/one-small-but-important-effort-to-make-twitter-safe-for-women/382484/

Morris, K. (2011, October). Reddit shuts down teen pics section. *The Daily Dot.* Retrieved 2016-09-01, from http://www.dailydot.com/society/reddit-r-jailbait-shutdown-controversy/

Morris, K. (2012, April). Reddit moderator banned for selling his influence. *The Daily Dot.* Retrieved 2016-08-30, from http://www.dailydot.com/society/reddit-hire-spam-ian-miles-cheong-sollnvictus/

Olanoff, D. (2015, August). Reddit Names Marty Weiner, Founding Engineer At Pinterest, Its First CTO. *TechCrunch.* Retrieved 2015-09-25, from http://techcrunch.com/2015/08/18/reddit-names-marty-weiner-founding-engineer-at-pinterest-its-first-cto/

Postigo, H. (2003). Emerging sources of labor on the Internet: The case of America Online volunteers. *International review of social History, 48*(S11), 205–223. Retrieved 2015-09-23, from http://journals.cambridge.org/abstract_S0020859003001329

Postigo, H. (2009, September). America Online volunteers. *International Journal of Cultural Studies, 12*(5), 451–469. Retrieved 2015-08-19, from http://ics.sagepub.com.libproxy.mit.edu/content/12/5/451

Preece, J., & Shneiderman, B. (2009). The reader-to-leader framework: Motivating technology-mediated social participation. *AIS Transactions on Human-Computer Interaction, 1*(1), 13–32. Retrieved 2015-01-25, from http://aisel.aisnet.org/thci/vol1/iss1/5/?utm_source=twitterfeed&utm_medium=twitter

Reddit downgrades technology community after censorship. (2014, April). *BBC News.* Retrieved 2016-09-01, from http://www.bbc.com/news/technology-27100773

Shaw, A., & Hill, B. M. (2014, April). Laboratories of Oligarchy? How the Iron Law Extends to Peer Production. *J Commun, 64*(2), 215–238. Retrieved 2015-07-02, from http://onlinelibrary.wiley.com/doi/10.1111/jcom.12082/abstract doi: 10.1111/jcom.12082

Sullivan, A. (2014, November). The SJWs Now Get To Police Speech On Twitter. *The Dish.* Retrieved 2016-02-04, from http://dish.andrewsullivan.com/2014/11/10/the-sjws-now-get-to-police-speech-on-twitter/

Terranova, T. (2000). Free labor: Producing culture for the digital economy. *Social text*, *18*(2), 33–58. Retrieved 2016-02-07, from http://muse.jhu.edu/journals/social_text/v018/18.2terranova.html

Zhu, H., Kraut, R. E., & Kittur, A. (2014). The impact of membership overlap on the survival of online communities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 281–290). ACM. Retrieved 2016-01-06, from http://dl.acm.org/citation.cfm?id=2557213

Chapter4

Community-Led Experiments in Platform Governance

In this chapter, I introduce CivilServant, novel software I created that online communities on reddit use to evaluate their governance practices, share the results, and replicate other communities' policy experiments. I describe five central design considerations for any effort to develop community-led experiments: community participation, research ethics, experiment validity, transparency, and deliberative replication.

The CivilServant project addresses these design challenges with what I call the *community knowledge spiral*, a process for conducting governance experiments with community input and oversight. I also describe the software architecture that I developed with Merry Mou to support this process and manage experiments. I evaluate the system by summarizing community responses to the experiments and reporting early uses of research findings by communities and by the platform's designers. These early findings offer evidence that social experiments can generate informative governance knowledge in ways that are accountable to the people they govern.

As social platforms and intelligent agents become routine parts of daily life for billions of people, the public has come to expect these systems to address deep-seated social ills. Platforms are currently asked to manage social problems including terrorism (Lomas, 2017; Wark, 2016), racial profiling by police (O'Donovan, 2016), suicide (Metz, 2017), self-harm (Chancellor, Pater, Clear, Gilbert, & De Choudhury, 2016), eating disorders (Chancellor et al., 2016), hate speech (D. K. Citron & Norton, 2011), child pornography (M. N. Thakor, 2016), human trafficking (Casteel, Thakor, Johnson, & others, 2011), misogyny (Banet-Weiser & Miltner, 2016), racial discrimination (Doleac & Stein, 2013; Edelman, Luca, & Svirsky, 2016), access to medical treatment (Tan, 2016), road safety (Martinez, 2015), copyright violation (Seltzer, 2010), and political polarization (Sunstein, 2009), to name a few. In recent years, mainstream advocacy organizations have established branches in Silicon Valley, hoping to influence U.S. platforms to adopt corporate policies on issues historically addressed through legislation (Chan, 2017). Even as platform operators attempt to retain positions of nonintervention (Gillespie, 2010), the designers and researchers who manage those systems have arguably become powerful civil servants who shape and govern 21st century human affairs (S. Geiger, 2015).

The pressure on platforms to become policymakers relies on an assumption that platforms possess effective means to govern society.[1] Yet well-intentioned interventions and design patterns have often increased a problem or caused disastrous side-effects for years before the effects were known (Cheng, Danescu-Niculescu-Mizil, & Leskovec, 2014; Halfaker, Geiger, Morgan, & Riedl, 2012). Furthermore, just as platform policy evaluation has not matched the rate of new interventions, the diversity of that research has not scaled to match the variations in human culture governed through online platforms (Hill & Shaw, n.d.). Among the research that is available, many findings remain secret within companies that are incentivized to protect their reputations and their intellectual property (Matias, 2016d). Consequently, public assumptions about the benefits of social interventions remain unproven assumptions while failures go unnoticed.

The stakeholders who wish to govern social behavior through platforms tend to choose between two strategies. The most common strategy requires

[1] Even if platforms place pressure based on a sense of their responsibility for what appears on their systems, this includes an unspoken assumption about the outcomes of removing that content.

platform operators to take governance power, build policy teams, and develop the research capacities to evaluate those policies (D. K. Citron & Norton, 2011; Gillespie, 2010; S. Geiger, 2015). Advocates and governments then attempt to hold platforms accountable for their use of governance power (MacKinnon, 2012). A second strategy delegates power from platforms to civil society organizations and volunteer moderators, who then create and enact their own local policies in formal relations with platforms or through community-created governance infrastructures (Postigo, 2009; Grimmelmann, 2015; D. Citron & Wittes, 2017; R. S. Geiger, 2016). Although the civic labor of delegated governance can be difficult to sustain (Matias, 2016a), this delegation strategy can scale governance work, adapt to cultural differences, and make public accountability a civic process rather than a corporate process (R. S. Geiger, 2016). Yet unlike platform operators, those who hold delegated power almost never have the capacity to evaluate the outcomes of their policy work to reduce hate speech, address child pornography, enforce copyright laws, or govern public discourse.

In this chapter, I introduce CivilServant, a novel system that online communities use to evaluate the outcomes of their social policies, share the results, and replicate other communities' interventions. Communities that work with CivilServant set the research goals, define the policies to be tested, and openly discuss fully-transparent results. The software collects data with community consent, coordinates interventions, generates results, publishes findings, and coordinates participant debriefings. Researchers facilitate discussions about evaluation design, configure studies, and participate in community debriefings about research findings.

CivilServant participates in a history of debates on the role of social experiments in democratic societies. I situate CivilServant within that history, offer design considerations for community-led randomized trial infrastructures, describe the system, and summarize the research process. I also report early findings from two large-scale policy evaluations by communities with over 12 million participants, evaluating the system as critical infrastructure and observing the early uses of research findings. I conclude by discussing the implications for a democratic, experimenting society where delegated governance power is evaluated independently by communities at scale.

Social Experiments in Democratic Societies

In *The Open Society and Its Enemies,* Karl Popper reflects on the uses of causal inference in social policy. Writing from New Zealand in exile from Nazi-controlled Austria, Popper describes social experiments in what he calls "open" and "closed" societies. In closed societies, paternalistic experts use the sciences to shape public behavior toward utopian goals, justifying their actions with the argument that "the learned should rule" (Popper, 1947, 107). In open societies, social experiments support the public to evaluate government policies "so that bad or incompetent rulers can be prevented from doing too much damage" (Popper, 1947, 107). Popper sees potential experiments everywhere, from the smallest action in the economy to broad policies that shape the direction of populations. Writing that the person "who opens a new shop, or who reserves a ticket for the theatre, is carrying out a kind of social experiment on a small scale," Popper imagines how large numbers of small experiments could benefit society (Popper, 1947, 143). For Popper, these experiments are more than a means of understanding behavior; they are political systems for social improvement through democratic rejection of ineffective policies and leaders.

If small-scale experiments were more common, writes Popper, "it might lead to the happy situation where politicians begin to look out for their own mistakes instead of trying to explain them away and to prove that they have always been right" (Popper, 1947, 143). Yet experimenters in a closed society, who Popper calls "utopian engineers," shape society without regard to the public's views:

> the Utopian engineer will have to be deaf to many complaints; in fact, it will be part of his business to suppress unreasonable objections. But with it, he must invariably suppress reasonable criticism also. (Popper, 1947, 140-41)

Fifteen years after Popper made these arguments, the methodologist and founding figure of policy evaluation Donald Campbell outlined a practical vision for social experiments in the governance of democratic societies. By 1971, the U.S. government was already converting recordkeeping to thousands of IBM 3/60 systems, imagining the use of data to improve education, fight poverty, and usher in a "Great Society" (Johnson, 1966; Oakley, 2000). As the U.S. government adopted research methods from Campbell's textbooks (Campbell &

Stanley, 1963), he worried that government policy experiments would threaten the "egalitarian and voluntaristic ideals" of democracy and lead to the "authoritarian, paternalistic imposition" of Popper's closed society (Campbell, 1998). Campbell argued that while ignorance of policy outcomes is a serious peril, it is also perilous to develop and use experimental knowledge apart from the democratic process.

In "The Experimenting Society," a lecture that policy evaluators photocopied and passed around for decades before it was published, Campbell outlined statistical and social processes for democratic field experiments. He proposed experiments where citizens are "co-agents directing their own society," defining goals, shaping variables, designing interventions, and actively interpreting, re-analyzing, and debating experiment results (Campbell & Stanley, 1963, 49). Campbell challenged methodologists to redesign their methods to include "individual participation and consent at all decision levels possible" (Campbell, 1998, 42). At a time when field experiments were rare, Campbell imagined a society where local communities conducted plentiful policy studies in a constellation of disputatious experimenters: "citizens not part of the governmental bureaucracy will have the means to communicate with their fellow citizens disagreements with official analyses and to propose alternative experiments" (Campbell, 1998).

By advocating for democratic networks of replication and cross-validation, Campbell anticipated later developments in feminist theory that grounded empirical research in the position and perspectives of the researchers, according to the feminist sociologist Anne Oakley (Oakley, 2000). In Campbell's original speech to the Russell Sage Foundation, he hoped that faculty at small regional U.S. colleges would be funded to conduct a multiplicity of new community-based experiments and replications with local government policy (Campbell, 1981). Instead, Campbell's proposal remained a thought experiment distributed and debated by practitioners within the policy evaluation field (Oakley, 2000).

With CivilServant, I am adapting the idea of an experimenting society to platform governance. By developing a system supporting plentiful, community-led policy experiments, I am working toward an open society, where the public gains the benefits of experimental knowledge with the benefits of a consequential voice on the policies that govern our lives. This chapter reports early findings toward those goals.

Delegated Governance Online

Platforms have delegated governance power to volunteers and civil society organizations since the earliest connected social technologies. In the 1980s, *conference hosts* on the WELL, BBS *SysOps*, and UseNet *moderators* created and enacted community policies (Rheingold, 1993; Bruckman, Curtis, Figallo, & Laurel, 1994). When for-profit companies were permitted to operate online in the 1990s, volunteer *community leaders* on AOL governed its many chatrooms (Postigo, 2009). Internet users continue to create and enact policy on all major internet platforms, including Wikipedia (Forte, Larco, & Bruckman, 2009), Facebook (Facebook, n.d.), Twitter (Matias et al., 2015; R. S. Geiger, 2016), reddit (Matias, 2016a), and Xbox (Good, 2013). Responsibilities for identifying and responding to copyright violations and child pornography are delegated to third-party corporations (Seltzer, 2010) and nonprofits (M. Thakor & others, 2013), an approach that some legal theorists have also suggested for hate speech (D. Citron & Wittes, 2017). The public accountability of these delegated authorities ranges widely, from democratic communities with transparency reporting and elections to organizations that operate in secret.

In the first deployment of CivilServant, I worked with volunteer moderators on reddit, a social news platform whose culture and system are well-suited to community-led experiments. On many platforms, independent randomized trials would attract legal risks related to Terms of Service and computer fraud regulations (Sandvig, Hamilton, Karahalios, & Langbort, 2014). On reddit however, independent data collection is routine to community moderation (Kiene, Monroy-Hernández, & Hill, 2016). A strike by over a thousand reddit communities in 2015 demonstrated moderators' appetite for participatory policy-making in community governance (Centivany & Glushko, 2016). This strike also revealed the network structure of moderator governance behavior across tens of thousands of communities (Matias, 2016b). Based on this research, I chose reddit as the research site because its communities already form a disputatious network of delegated governance power.

Related Systems

Systems supporting randomized trials are now a common component of social technologies, and experimentation is routinely carried out by software engi-

neers and designers. Within Microsoft, a causal inference team supports hundreds of concurrent experiments per day on the Bing search engine (Kohavi et al., 2013). At Facebook, the open source PlanOut system supports engineers to incorporate a range of randomized trial designs into product testing (Bakshy, Eckles, & Bernstein, 2014). Companies including Optimizely provide randomized trials as a service to third parties; a single page of the New York Times often includes at least half a dozen A/B tests of headlines and other design features (Reisman, 2016). All of these teams and systems provide sole experimenting capacity to the organizations that control the design and use of platforms. With CivilServant, I have built a system that supports platform users to conduct their own research independently from platforms, including research that tests users' influence on the platform operator's systems.

Systems like CivilServant collect data and coordinate users within a larger platform, creating alternative knowledge for community mutual aid. Creators of these "successor systems" often attempt to restructure the power relations encoded into a larger system through software, data, and collective action, according to Stuart Geiger (R. S. Geiger, 2014). For example, with Turkopticon, digital laborers share mutual-aid information to evaluate the people who offer them work (L. C. Irani & Silberman, 2013). Third-party audits of platform governance and community moderation technologies serve similar functions by providing independent knowledge and supporting community processes that restructure public life online, according to Geiger (R. S. Geiger, 2016; Matias et al., 2015). These successor systems often piggyback on those larger systems for core functions (Grevet & Gilbert, 2015). Since social change is often a primary design goal of these systems, they often function as critical infrastructure, generating ongoing knowledge that continues to serve community needs (L. Irani & Silberman, 2014). CivilServant relates to these mutual aid, critical infrastructures by supporting communities to evaluate their effects.

As a socio-technical system, CivilServant is also related to other social infrastructures of experimental knowledge-making. Organizations including the Cochrane Collaboration, Campbell Collaboration, and Evidence in Governance and Politics[2] facilitate peer support for experimenters and publish meta-analyses of findings across health, social policy, and international development (Bero & Rennie, 1995; Chalmers, Hedges, & Cooper, 2002). In the UK, the What

[2]http://egap.org

Works Network supports local government to design and conduct novel studies and replications of social policy experiments (Alexander & Letwin, 2013). Like these initiatives, the CivilServant project publishes findings and literature reviews on its website while also organizing communities to develop new, usable evidence on the effects of platform-based policy interventions.

Design Considerations

Community Participation

Any process for evaluating social interventions will structure the power held by the intervention's stakeholders in some way. Among traditions of participatory evaluation, some movements prioritize close collaboration with existing power structures while others prioritize direct work with those who hold the least power (Cousins & Whitmore, 1998). Platform governance brings together a complex network of actors across communities, preventing platform power from being so easily classified. For example, moderation often involves exercising power in multi-party conflict situations to protect and govern a community that may include tens of millions of people (Keegan & Matias, 2015). While platforms delegate higher levels of power to moderators than many other users, those moderators also have relatively little power and agency compared to platforms and law enforcement. Furthermore, some of the least empowered people in platform governance are those who allegedly organize to harm others. No process can protect the most vulnerable while guaranteeing equal participation of all stakeholders in the evaluation of moderation work.

I attempted to design the CivilServant system and its research processes to expand the potential participation of anyone involved in or affected by online moderation. This design work draws inspiration from Arnstein (Arnstein, 1969), who imagines a ladder of citizen participation on a scale from non-participation to tokenism to citizen power. While the project cannot and should not offer equal power to all stakeholders, it can increase the level of participation offered to anyone involved. By supporting communities to conduct their own studies, publishing the results transparently, and requiring communities to openly discuss the results, CivilServant expands participation for all stakeholders, including those who are judged to violate community policies. Communities and others affected by moderation actions become informed and in

some cases consulted through public discussions of findings. Moderators who already hold delegated power from platforms gain new information for their discussions on the uses of that power.

Research Ethics

Academic and professional social computing researchers are currently re-negotiating research ethics practices after highly-publicized controversies over corporate and university experiments conducted without participant consent or ethics board review (Grimmelmann, 2015; Fiesler, Chancellor, Hoffmann, Pater, & Proferes, 2016). In the U.S., many of these conversations struggle with the legacy of institutional infrastructures developed to govern medical research, models of research ethics that poorly address the range of interventions and risks introduced by contemporary online social research (Vitak, Shilton, & Ashktorab, 2016).

I take the view that practical attempts at social change offer powerful opportunities to produce usable knowledge through participant-led methods (?, ?). Given the tremendous power exercised by those who govern social behavior online, I also believe that power-holders have an obligation to evaluate the outcomes of their governance work (Meyer, 2015). Kurt Lewin, a pioneer of participatory field experiments in factories and education, reportedly described this relationship with the slogan "no action without research; no research without action" (Adelman, 1993).

Large-scale governance experiments also entail complex relations of risk and harm. Because this research focuseses on governance, I have drawn inspiration from recent conversations on experiment ethics in political science, where multi-party interests and public goods often conflict in complex ways (Desposato, 2015). In political science, experiment outcomes can include direct and secondary effects on people's rights, beliefs, representation, and well-being. To manage these risks, political scientists are developing novel methods for consent from groups, stakeholders, and participant representatives, as well as testing novel debriefing procedures (Desposato, 2014). The CivilServant research process currently supports community discussion in experiment planning, individual debriefing, and community debriefings. As a growing diversity of communities use CivilServant to test their governance work, I plan to adapt the system to support empirical research on novel research ethics procedures.

Internal and External Validity

The most valuable experiments generate useful findings that also contribute to generalizable knowledge. Within the social sciences, the quality of research procedures and an experiment's applicability to the population being studied are called internal validity. Discussions of external validity are concerned with the wider applicability of study results and related theories; with more replications, a finding is often considered more externally-valid (Campbell & Stanley, 1963).

While designing CivilServant, I faced an apparent trade-off between internal and external validity. Because replications are rare in social computing research, the emerging field struggles with problems of external validity (Hornbæk, Sander, Bargas-Avila, & Grue Simonsen, 2014). If I created a simple tool that required no involvement from expert researchers, I might be able to grow external validity by scaling the number of replications studies more quickly. With a semi-bespoke approach, I could improve the internal validity of a smaller number of early studies while building software processes that could be used in future replications. Regression adjustment variables improve the precision of experiment estimates; interference monitors adjust for treatment overspill; power analysis systems guide study duration calculations; intervention procedures support assignment by discussion or individual; and block randomization over time protects the study from being spoiled by software errors (Gerber & Green, 2012). As the demand for community experiments grows, I expect that these early investments in internal validity will strengthen the external validity that may come from larger numbers of community replications.

Open Knowledge and Transparency

I created CivilServant to generate open knowledge on the results of platform-independent causal research online. Since the primary audience for that knowledge includes community members who may just be starting to develop their data literacy, CivilServant prioritizes general-audience publishing and the community engagement needed to reach communities with findings. By publishing full software and analysis, CivilServant can contribute to an open research culture among other scholars who can query and replicate experimental findings (Nosek et al., 2015).

While designing CivilServant, I also encountered an apparent tension be-

tween research openness and privacy that I have yet to resolve. If CivilServant openly publishes the full datasets from studies, scholars and participants could confirm, question, or extend new findings by re-analyzing the collected data. While CivilSrvant might increase the trust of other scholars by publishing experiment data, I am still assessing the risks to participants and communities of releasing experiment data. For example, if the contents of moderated messages were made public, the people who made those comments could be targeted by others, or the people they harassed could become targets a second time out of retaliation for the act of moderation. For now, CivilServant keeps all experiment data private. I may consider releasing data from future studies if communities can agree on processes that reconcile research transparency with participant privacy.

Deliberative Replication in an Experimenting Society

In Campbell's imagined society, field experiments are a plentiful form of knowledge generated by the public rather than a rare resource controlled by experts (Campbell, 1998). In such circumstances, many of the values and power structures typically associated with experimentation could become transformed. To pick one example, the apparent tension between privacy and re-analysis might become less prominent as replications become easier and more common. Furthermore, longstanding arguments that experimental knowledge can be used to override citizen interests may become transformed if disputatious networks of community experimenters are able to bring their own findings into decision-making processes. I can only speculate on the outcomes of an experimenting society. Yet by drawing from traditions of participatory research and favoring widespread deliberation, I hope to reduce the chance of creating worse outcomes for society (Bardzell, 2014).

With CivilServant, I use design and social processes to support delegated platform governance through plentiful, deliberative social experiments. As communities conduct studies, each iteration of the open source software reduces the work required of other communities to conduct new studies and replications. I also invest substantial effort into community outreach and education to spread the skills needed to meaningfully discuss findings and design new studies.

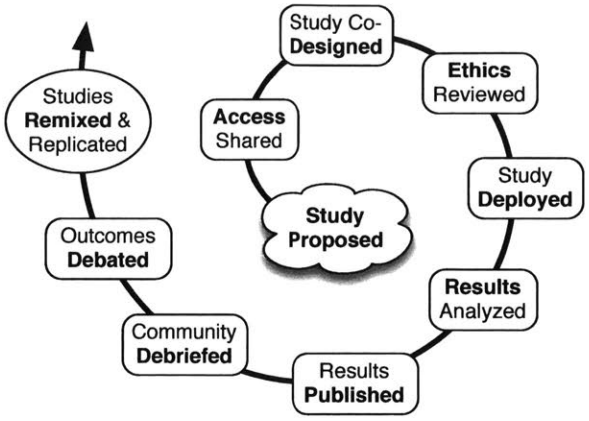


Figure 4-1: **The Community Knowledge Spiral**: CivilServant studies use processes that grow community-led experimentation.

The CivilServant System

Designing Studies with CivilServant

Studies conducted with CivilServant follow a social and technical process that I call the community knowledge spiral (Figure 4-1). The spiral starts with community interest, continues through the deployment and interpretation of a study, and continues to grow through replications by other communities.

Community Interest

The process of using CivilServant begins when an online community identifies a testable question about the effects of moderation work, usually in conversation with the CivilServant project. Because many communities on reddit already collect and process data, they often possess substantial knowledge about the kinds of interventions and measures that would be feasible in a field experiment. The CivilServant study designer mainly contributes to conversations by facilitating a discussion of the mechanisms for delivering randomized interventions that could be coordinated by the software and compatible with community goals.

Permission for Data Access

If moderators wish to continue, their next step is to add the CivilServant bot to become a moderator with archival-only privileges to their community. At this

point, the system collects historical and ongoing data on submitted posts, comments, the behavior of ranking algorithms, moderator activity, and in some cases, data from third party systems that moderators use to coordinate their actions. As an agent with archival-level access, the CivilServant system can process data without the ability to take action on behalf of the community. If moderators wish to end the relationship and prevent further data collection, they can revoke the system's access at any time.

Study Design, Power Analyses, and Pre-Analysis Plans

After CivilServant has collected data for a period of time, I use the data to construct a formal study design from the community requests. The study design takes the form of a pre-analysis plan and a power analysis report. For studies that replicate or remix variables from previous studies, I produce the pre-analysis plan using a set of R Markdown and LaTeXdocuments that take CivilServant data as input. At present, each novel study design requires a researcher to write a short literature review and experiment design, creating the templates that can be used to fill other replications.

The pre-analysis plan provides a community with a formal description in non-expert language that explains study's goals, intervention, variables, and analysis procedures. Further conversations about the details of the study center around all aspects of this document, which becomes a running record of community decisions about the intended study. Throughout this process, reports of power analyses inform the community about their chance of observing certain effect sizes in the time-spans they are interested to conduct the study. While no community has yet rejected a study proposal after seeing how long it might take, the power analysis helps communities set the timeline for an experiment and understand the value of waiting for the answer. All pre-analysis plans are published to the CivilServant repository on the Open Science Framework service, keeping the identity of participating communities private.

Ethics Approval

Once communities decide the area of policy and kinds of measures used in a study, I consult its compatibility with the CivilServant project's existing IRB permission. The project operates under three kinds of IRB agreements for data collection, routine policy evaluations, and higher risk studies.

Data collection of public information for the purposes of designing a study is covered under an observation-only IRB agreement. A second IRB covers a class of possible studies that involve routine moderation actions with minimal risks to participants in low-risk communities. This IRB excludes communities that trade resources, offer advice and mental health support, or engage in conflict with other communities. The IRB, which waives consent requirements but requires communities to be debriefed, also excludes studies involving banning accounts and related interventions carrying social costs that are not easily reversible. Most studies conducted by CivilServant fit within this IRB for routine policy evaluation. I have also sought and received IRB approval on a per-study basis when the risks are higher, or when individual consent and debriefing are appropriate. At MIT, this process typically takes less than a month and occurs in parallel with the work of finalizing the study design with communities.

Finalizing Study Designs

When moderators and subreddit communities discuss the pre-analysis plan, they often notice details of the study that need to be adjusted, as well as changes to their own moderation infrastructure that may be required. Participants often think of potentially-confounding factors on the platform or within their community, factors that are subsequently accounted for in regression adjustment variables or the design of the experiment. For example, in one study, participants wondered if the outcomes might be affected if reddit's algorithms promoted some conversations beyond their community. I added this factor as a regression adjustment variable.

Studies involving new interventions or variables sometimes require new software development. With replications and remixes of features from past studies, the software records the final study design using a domain specific language similar to Facebook's PlanOut (Bakshy et al., 2014).

Recipes for Theory-Informed Interventions

In most studies I attempt to bridge between theories of human behavior and the interventions that communities test, while reserving community agency over the intervention that they choose. For example, all of the studies designed with the CivilServant project allow for variations in the messages that accompany an intervention. Theories from social psychology offer guidance on the possible effects of different styles and claims within those messages. In study design facilitation, I have supported communities to adopt theoretically-informed language by listing the categories of language that theory might suggest. For example, in statements of rules, a community might wish to make appeals to widely-held norms, to authority, or to the consequences of an action. Using an online collaborative text editor, I provide communities with a list of "ingredients," and invite them to fill out the details with messaging alternatives that they can decide on as a group.

Testing & Deploying Community Experiments

Once a community's moderators agree to a final experiment design, I submit the final pre-analysis plan to the Open Science Framework and test the live experiment software for compatibility with the community's other systems. Once all issues are resolved, I generate documented randomizations and deploy the experiment for the agreed-on period, monitoring experiment activity for consistency and compliance to the study procedures. All studies are block-randomized; if software or compliance errors occur, small, individual randomization blocks may be removed without spoiling the balance of the sample.

Concluding Community Experiments

As the study proceeds, I regularly notify communities about the rate of progress toward the agreed number of randomizations. Some designs include a stop rule for ending the experiment early if large or harmful effects have been observed. Upon reaching the stop rule, I generate results and notify the community if the stop rule criteria have been met. The community can then decide if the study should continue for its full duration.

Analyzing Findings

At the conclusion of a study, the CivilServant software generates dataframes for each of the hypotheses in the study. In a process that is not yet fully automated, I produce a technical experiment report using R Markdown templates that apply the statistical methods listed in the pre-analysis plan and present them alongside summary statistics. For study remixes and replications, the work of confirming the validity of data collected for each experiment block remains the most bespoke part of the process. This involved reviewing the software logs and summary statistics to identify any errors in data collection.

Community Debriefings

The CivilServant project holds a policy of maximum practical disclosure to involved communities and participants. The software is able to send a debriefing message to every participant whose activity was included in any dependent or adjustment variable in the study. All study results are public, and CivilServant requires that all participating communities agree to host a "community debrief," a public conversation within their community to report and discuss the results. To support that conversation, I publish a public-audience summary of the study motivations, procedures, and findings on the CivilServant website. I also offer to participate in the debriefing conversation and answer questions about the findings. These conversations tend to be the first of a community's discussions and debates of what the findings mean for their governance practices. I also notify the platform operators, often for the first time, that the community has conducted and completed a new study with us.

Remixing and Replicating Studies

I chose to deploy CivilServant on reddit because it hosts many different online communities. As results of each study are made public, other communities can choose to replicate each the study or remix its features. Community replications may strengthen the external validity of research findings while also discovering variations among groups that predict differences in experiment outcomes (Hill & Shaw, n.d.). At the time of writing, two groups of subreddits are independently considering replicating each others' studies in parallel: one group of four communities and another group of two communities.

System Architecture

I designed the CivilServant system to aid in study design, collect relevant data, manage interventions, and support processes for analysis, reporting, and debriefing. Implemented together with Merry Mou[3] in Python, R, and MySQL, the system was managing hundreds of millions of records across a distributed hardware infrastructure at the time of writing (Figure 4-2).

The analysis infrastructure includes software for scoping problems, designing studies, generating dataframes, conducting statistical analyses, debriefing participants, and removing them from studies upon request. We conduct data cleaning and validation in Jupyter Notebooks and publish them as part of documenting experiment procedures. We generate all experiment results and reports using R Markdown, as specified in a pre-analysis plans. The CivilServant reporting repositories are linked with the Open Science Framework, which also publishes all pre-analysis plans (Columns 1 and 4 in Figure 4-2) (Miguel et al., 2014; Nosek et al., 2015).

We developed the data collection and intervention architecture to offer abstracted experimentation capacities across multiple platforms. Individual study designs are configured with a domain-specific language that describes the platform, community, authentication details, intervention arms, conditional logic, and randomizations for a study. A job scheduler that monitors API keys and rate limits manages requests for data and interventions across a pool of authenticated user accounts associated with CivilServant. At the time of writing, the intervention architecture includes connections to reddit and to several third-party moderation systems used by moderators (Columns 2 and 3 in Figure 4-2).

Community Experiments with CivilServant

Increasing Newcomer Policy Compliance

Moderators of r/science, a 13-million subscriber community at the time, approached CivilServant in February 2016 to conduct the first CivilServant study, which we carried out from August 25, 2016 to September 23, 2106. In this community, over 1,200 volunteer university faculty, graduate students, and undergraduates organized to foster large-scale discussions of new peer reviewed

[3] I use "we" in the system architecture section to acknowledge Merry's considerable contributions to the CivilServant project.

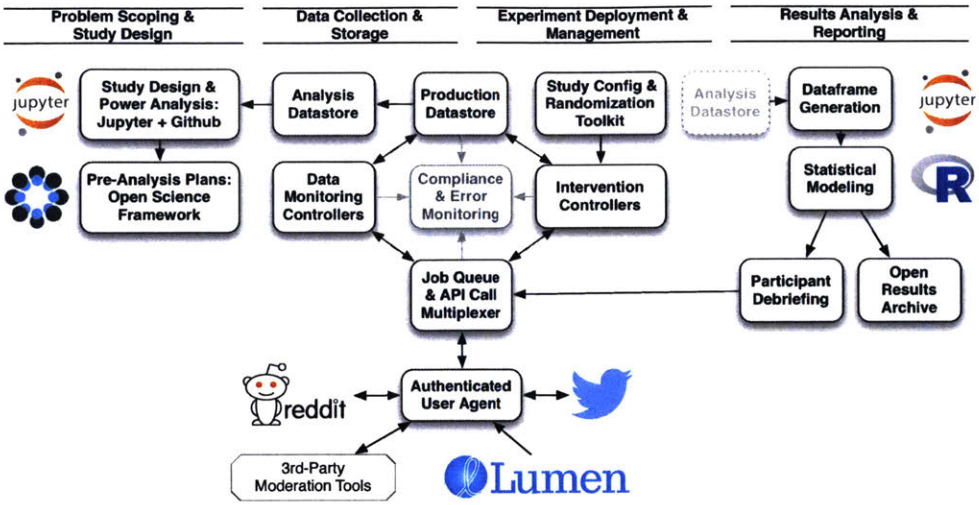


Figure 4-2: System architecture of the CivilServant system, grouped by the function of each system component in the process of designing, conducting, and interpreting a field experiment.

research. Community policies include expectations that commenters focus on the discussion topic, avoid abusive language, avoid giving medical advice or personal anecdotes, limit jokes, and cite peer reviewed research when criticizing established scientific theories. Prior to the study, moderators removed over 1,200 comments per day on average, across an average of 147 discussions per day. Of the comments they removed, 39% came from first-time commenters.

In this study, which I describe in detail in chapter five, moderators crafted messages explaining community norms that could be posted to the top of discussion threads. They used CivilServant to randomly assign those messages to some discussions after they appeared and not to others. Across 2,190 discussions of academic publications and 24 live question-answer sessions with academic researchers, we found together that in discussions that omitted posting the rules, a first-time commenter has a 75.2% chance of complying with community norms, and that posting the rules has a positive 7.3 percentage point effect on the chance that a newcomer's comment will be accepted by moderators, on average in the community. Although we expected that posting the rules more visibly would deter newcomers from participating, we also found that posting the rules increases the incidence rate of newcomer comments by 38.1% on average. Overall, posting rules could prevent over 1,800 first-time commenters from engaging in unacceptable behavior and attract over 9,600 new commenters per month on average (Matias, 2016c).

We held a community debriefing with the science community in October 2016, a conversation that included 478 comments, attracted over 14,000 votes, and was viewed by over 240,000 readers. Other communities on reddit also discussed the findings widely. In the months that followed, the record of our debriefing discussion was occasionally referred to by other communities as they decided how to govern their subreddits. The following spring, the reddit platform began testing new features to support communities across the site to make their rules more visible to participants.

Governing Human & Machine Responses to Unreliable News

Moderators of the r/worldnews community approached me in October 2016 to test methods for governing the reception and spread of unreliable news in their subreddit. Articles from unreliable sources were 2.3% of all submissions to the community, which reviewed an average of 450 articles per day. Moderators wished to avoid banning certain news websites but also wished to encourage reader skepticism toward those sources. Together, we were also concerned that increased fact-checking activity might influence the behavior of reddit's recommendation systems, potentially causing fact-checked articles to be promoted more widely by those algorithms.

In this multi-armed study design, which I describe in greater detail in chapter six, moderators posted messages encouraging readers to fact-check articles with links to further evidence. In a second message, moderators added language encouraging readers to use reddit's voting systems to dampen the algorithmic spread of unsubstantiated articles. The CivilServant system randomly assigned these messages across 1,104 posts from December 7th 2016 to February 15, 2017. The system observed the contents of comments, the algorithm "score" of each post every four minutes, and reddit's popularity rankings every four minutes. All arms that encouraged fact-checking increased the chance that individual comments and discussions would include links to further evidence (Matias, 2017). However, while the arm encouraging fact-checking caused the algorithmic ranking of news articles to reduce by a maximum of four positions over time in the top 100 entries on average. I failed to reject the null hypothesis for the arm that also encouraged voting. The findings confirmed our expectation that encouraging fact-checking would influence reddit's rankings, but the outcome was the opposite from our expectations.

We debriefed the r/worldnews community in a day-long public discussion on February 2017 that included 280 responses and received over 2,000 votes. Our results were discussed widely in other reddit communities. We also received over a dozen personal notes from r/worldnews participants about the study.

Studies In Progress

Since these first two studies, CivilServant is working with further communities on a wide range of upcoming experiments. Two subreddits are currently voting on studies that test reductions in recidivism rates among participants who are re-integrated into the communities after being banned. Another four communities are considering replications of the experiment first conducted by r/science. Further studies under discussion include efforts at conflict resolution in polarized discussion groups and interventions to mitigate the effects of automated copyright enforcement systems.

Evaluating CivilServant

Because policy evaluation is a multi-part process that often begins and ends with group decisions, researchers have struggled to compare the policy contributions of different experimenting approaches (Contandriopoulos, Lemire, Denis, & Tremblay, 2010). In computer science, systems papers about experimentation infrastructures tend not to evaluate the social outcomes of experimental knowledge, preferring instead to report design considerations, implementation details, and discussions on the role of experiments in computer science (Kohavi et al., 2013; Bakshy et al., 2014). Critical infrastructure systems such as Turkopticon have typically been evaluated for their role to foster critical thinking among those who learn about the system or use it, as is common in critical design (R. S. Geiger, 2014; L. C. Irani & Silberman, 2013; Dimond, Dye, Larose, & Bruckman, 2013; L. Irani & Silberman, 2014). The field of policy evaluation tends to evaluate an experimentation approach by the values and uses of research. Practices that successfully implement their stated values and increase the adoption of research tend to be favored (Cousins & Whitmore, 1998). Yet because research adoption is a complex social process, literature reviews have failed to find relationships between the uses of research and the

details of experimentation practices (Contandriopoulos et al., 2010).

To evaluate CivilServant in this early stage, I take approaches from system design, critical infrastructure, and policy evaluation. As in other systems papers about experimentation infrastructures, I have reported the primary design considerations and the details of the system implementation as it was used in two large-scale community experiments on reddit. Because I designed CivilServant as critical infrastructure, I report qualitative findings on the kinds of critical perspectives that over a thousand participants brought to community debriefings.[4] Finally, I evaluate CivilServant as a process for community policy evaluation, reporting early results on the use of community-led experimental knowledge by communities and the reddit platform. Findings on the use of study results are informed by interviews with over a dozen moderators, participant observation in text conversations with over a dozen subreddits, and emails exchanged with the reddit platform employees. Subreddits, which ranged from thousands of subscribers to tens of millions, were included in the sample if they conducted an experiment, discussed the experiment results in public, or if moderators reported that they discussed the results privately. Moderators were sampled from communities that conducted or discussed experiment results, including communities that adopted the evaluated policies and communities that did not adopt the policies.

Community Debriefings

I observed community responses across more than a thousand public and private responses in two community debriefings. Communities held these debriefings by posting the results in an open discussion thread in the subreddit and "pinning" the discussion to the top-most recommended conversation for at least one full day. I also made myself available to answer questions during those debriefings.

In debriefing discussions, many commenters recalled encountering the intervention and shared personal stories that related to the findings. For example, one person in the news study experiment reflected: "I focus more on reading comments than the article itself. If people are fact-checking the article in the comments, I assume most will see it." Personal stories often open longer discussions about the purpose and legitimacy of moderation policies. One com-

[4] I have obfuscated all quotations from these debriefings.

menter reflected that "After I start typing, I see that a rule that conflicts with my comment and curse." Someone else replied "Isn't that the point?" and asked if the original commenter considered the outcome beneficial or not.

Comments sometimes offered direct critiques of community policies. For example, some commenters argued that the science discussion community should permit and support contributions from climate change skeptics. Others argued that moderators should be more flexible, allowing more jokes and "cool shit" alongside peer reviewed research. In the news discussion community, commenters argued that moderators should have included state-sponsored media in the fact-checking intervention. These criticisms sometimes prompted extended community discussion about preferred approaches to moderation. When one news commenter complained that encouragements to fact-check amounted to telling readers how to think, other commenters responded that the study was encouraging critical thinking and greater intellectual independence among readers.

Many commenters shared questions about research methodology in community debriefings. They asked questions about statistical significance, randomization methods, the choice of dependent variables, and confounding factors. Some suggested additional measures and hypotheses that could bring clarity to the findings. While I expected that I would need to explain more about my research methods, many statistics questions were answered instead by other community members. The demographics of reddit may explain why community members answered many of the statistical questions. On average in the United Sates, 82% of reddit users have some college education, twenty-three percentage points more than the rest of the population. The disparity is even higher among reddit users who browse the site for news (Barthel, Stocking, Holcomb, & Mitchell, 2016). Among subreddits with millions of subscribers, many of who have received a college education, it is common that communities already include active participants with statistical knowledge. The CivilServant project may encounter greater challenges with data literacy as it becomes used more widely.

During the debriefings, commenters also offered personal theories to explain experiment results. Many wondered if effects would endure as an intervention became less novel. Others reflected on details in the design of the reddit platform and the experiment that might have contributed to the results. Some shared stories about what they had learned from public-audience psychology

and sociology books.

Many on reddit participate in more than one community and wondered if interventions tested in one community might be useful elsewhere. In one case, over 15% of all comments in a community debriefing focused on the possibility of implementing the evaluated policy in a separate community. Other comments imagined the potentially beneficial or catastrophic effects of attempting the same intervention elsewhere. Some argued that we should have withheld sharing the results until completing further replications.

Commenters in both debriefings discussed research ethics. Several community members argued that I should release full datasets, leading to extended discussions of CivilServant's policies on privacy and anonymity. Others questioned the research ethics of the community interventions. "Did you do what Facebook did?" asked one participant, referring to a 2014 study that received widespread popular disapproval (Grimmelmann, 2015). In the discussion that followed, arguments over research ethics were interleaved with arguments over community policies. One participant appeared to argue that since the community's policies against abusive speech and personal attacks were an unjust form of censorship, then experimental interventions that reduce the rate of abusive speech also introduce serious harms that make the research unethical. These comments prompted extended discussions among the community about the justice of community policies and research ethics regulations in the United States. Although CivilServant offered participants the option to report any harms they experienced or retro-actively opt out of CivilServant studies, no one has yet contacted the project to do so at the time of writing.

Many people shared expressions of gratitude in community debriefings. In private and public messages, moderators and I were surprised at the frequency of thank-you messages. Several messages told a personal story, connected that story with concerns about broader trends in society, and thanked us for adding evidence to community governance. When moderators and I shared the results of our fact-checking study less than two weeks after the 2017 U.S. presidential inauguration, we expected that some U.S. commenters would interpret our work as politically-partisan. In community debriefings and private messages, people of all political affiliations thanked us and moderators for adding evidence into a conversation they saw as dominated by "bias" and "bullshit."

Some moderators expressed surprise at what they perceived to be a lack of substantial community criticism. Many expected that community debriefings

would attract complaints. Others disagreed. One moderator, who joined after reading a the community's moderation transparency report, saw the study as another example of moderator responsiveness to the community: "To me that indicated that the mods were really thinking about the readers." In r/worldnews, one moderator did anticipate that the findings would be popular, since readers frequently complained about tabloids and often replied to tabloid articles with profanity-filled complaints. Because this moderator saw the experiment as an effort to respond to community pressure, they expected the community to welcome the findings.

While community debriefing discussions were requested by hundreds of thousands of people and attracted meaningful commentary, CivilServant cannot observe what proportion of experiment participants they represent. CivilServant is able to observe commenting behavior, but the reddit platform does not provide third parties with the ability to track the viewing or voting behavior of platform viewers. Beyond participants that added visible comments, the system cannot not know the full number of silent viewers who observed experiment interventions and cannot match those participants with debriefing viewers. As with any public consultation, vocal participants included a small fraction of likely participants. In future studies, I hope to reconcile the values of widespread public debriefing with the limitations of platforms for managing very large discussions (Zhang, Verou, & Karger, 2017), and strong community norms against "message spam".

Uses of Community Experiment Findings

While the first findings from CivilServant were only published six months ago, my qualitative research on the uses of CivilServant results provides an early perspective on the project's outcomes. In the field of policy evaluation, where causal knowledge constitutes only one resource available to decision-makers, groups rarely adopt an intervention tested in randomized trials (Contandriopoulos et al., 2010). Research might become available after policymakers make a decision or might remain unread until external factors force a policy decision. Policymakers often read social research as "enlightenment" rather a judgment on the effectiveness of a specific intervention (Weiss, 1977). Resource limitations and political factors may lead policymakers to delay or set aside evidence-based policy ideas. Yet research read for general enlightenment can, in time, inform

those external forces as well (Weiss, 1979). In moderator interviews, content analysis of subreddit discussions, and correspondence with reddit employees, I found that communities' uses of CivilServant findings follow many of the usage patterns explored in the policy evaluation literature.

How Experimenting Communities Used Results

At the time of writing, none of the communities that conducted studies had changed their moderating practices after learning of results. Three months after I reported results to one community, a moderator explained that they intended to make changes, but more pressing demands had prevented them from finding the time to reconfigure their complex automated moderation system. In interviews, moderators of another community described hopes of a future decision to adopt the evaluated policy. One expected a substantial debate over the details of the policy. By demonstrating the effect of the intervention, CivilServant had opened up a complex decision where moderators might struggle to reach agreement.

How Other Communities Used Results

Moderators of communities beyond the ones that used CivilServant also read the results. Some of these moderators used the results to advocate for change, defend existing policies, and adopt personal moderation practices. In interviews, several moderators reported suggesting that their community adopt practices tested in one of the CivilServant studies. In one case, a community constructed their policy intervention to link directly to the document reporting study results. In other communities, after some moderators expressed skepticism that research findings would apply to their subreddits, the moderation team contacted CivilServant to conduct study replications. One subreddit had already been posting participation rules at the top of each discussion when CivilServant results were shared elsewhere. As conversations in this community became more contentious throughout 2016, moderators considered removing the message. In interviews, they reported that they discussed our experimental evidence in the conversation where they chose to retain rule postings.

In communities with less formal policy decision-making, research also influenced moderators' personal practices. In interviews, several moderators de-

scribed the ways that research results had led them to reflect on and change their personal moderation work. When moderators of one gaming community with over half a million subscribers read the r/science experiment results, they considered automating comments with the rules. When a group decision on an automated system did not materialize, individual moderators decided to personally-post messages with the rules on a case-by-case basis. In other communities, moderators who advocated for a policy were sometimes encouraged to try a practice for themselves before others adopt the idea. In these communities, experimental evidence can enlighten the work of individual moderators who pioneer and spread moderation practices to others.

How the reddit Platform Used Results

While I designed CivilServant to create platform-independent research, community interests often align with the interests of platforms. After debriefing r/science and r/worldnews about their findings, I notified the company that we had completed new research. In personal correspondence, employees described keeping our findings in mind when designing and testing new features across the platform. At the time of writing, the reddit platform is completing a randomized trial that tests the effects of a new feature to display community norms to newcomers who comment. [5] Employees described taking advantage of their greater control of the platform software to deploy a more nuanced randomization strategy, generating more precise estimates across different communities than our study permitted.

Findings

As the public turns to platforms to govern behavior and address enduring social problems, the public also needs methods to evaluate platform policies as part of an open society. We developed CivilServant to support communities to evaluate their uses of the power that platforms delegate to them. By contributing to a community knowledge spiral, communities add experimentation to their existing ways to evaluate policy, sharing findings and participating in conversations about the implications for their communities.

[5] `https://www.reddit.com/live/x3ckzbsj6myw/updates/71570f82-0a99-11e7-918d-0ee3534f4960`

In this chapter, I have reported design considerations, the public discourse in community debriefings, and the uses of community-led experimental knowledge. In two studies with CivilServant, communities tested the effects of interventions on the humans and algorithm behavior. In community debriefings, participants shared wide-ranging comments about governance policies, research methodologies, theories of human behavior, and research ethics. Communities were slow to implement the policies they evaluated, a pattern also observed in government policy evaluation. Yet the research findings informed individual moderator practices, community policies, and replications by the platform.

When designing the CivilServant software and social processes, I faced decisions about community participation, research ethics, methodology, and privacy that communities are likely to amend as the system is used more widely. Principled, thoughtful people may argue on competing sides for greater data transparency, privacy protections, or consent processes. I see these debates as evidence of the kind of open society we wish to foster. Rather than converge a notion of the best design, I have discovered the benefits of flexible architectures that prompt community discussion on important questions by offering communities the power to choose among options while protecting vulnerable participants. As CivilServant grows, I expect that these design decisions will more closely resemble the project's IRB process: defining areas of community deliberation and choice that remain constrained by the project's ethical, political, and methodological commitments.

Based on these early findings, I am hopeful that online communities, if they wish, could develop the network of deliberative experimenters imagined by Campbell in "The Experimenting Society." If online platforms are to become conduits for efforts to address much of humanity's fundamental social problems, society will need systems of knowledge production that dramatically scale and localize policy evaluation without restricting human autonomy and rights. A disputatious network of independent community policy evaluators might collectively develop and replicate effective governance practices at those scales. Because delegated power can be misused unjustly, community-led evaluation infrastructures like CivilServant may also offer the public valuable information for disputing ineffective or unjust uses of governance power on platforms.

References

Adelman, C. (1993). Kurt Lewin and the origins of action research. *Educational action research, 1*(1), 7–24. Retrieved 2017-03-29, from http://www.tandfonline.com/doi/pdf/10.1080/0965079930010102

Alexander, D., & Letwin, O. (2013). What Works: evidence centres for social policy. Retrieved 2017-03-26, from http://dera.ioe.ac.uk/17396/1/What_Works_publication.pdf

Arnstein, S. R. (1969). A ladder of citizen participation. *Journal of the American Institute of planners, 35*(4), 216–224. Retrieved 2017-03-26, from http://www.tandfonline.com/doi/abs/10.1080/01944366908977225

Bakshy, E., Eckles, D., & Bernstein, M. S. (2014). Designing and Deploying Online Field Experiments. In *Proceedings of the 23rd International Conference on World Wide Web* (pp. 283–292). New York, NY, USA: ACM. Retrieved 2015-10-03, from http://doi.acm.org/10.1145/2566486.2567967 doi: 10.1145/2566486.2567967

Banet-Weiser, S., & Miltner, K. M. (2016). # MasculinitySoFragile: culture, structure, and networked misogyny. *Feminist Media Studies, 16*(1), 171–174. Retrieved 2017-03-29, from http://www.tandfonline.com/doi/full/10.1080/14680777.2016.1120490

Bardzell, S. (2014). Utopias of Participation: Design, Criticality, and Emancipation. In *Proceedings of the 13th Participatory Design Conference: Short Papers, Industry Cases, Workshop Descriptions, Doctoral Consortium Papers, and Keynote Abstracts - Volume 2* (pp. 189–190). New York, NY, USA: ACM. Retrieved 2017-03-26, from http://doi.acm.org/10.1145/2662155.2662213 doi: 10.1145/2662155.2662213

Barthel, M., Stocking, G., Holcomb, J., & Mitchell, A. (2016, February). *Reddit news users more likely to be male, young and digital in their news preferences* (Tech. Rep.). Pew Research Center. Retrieved 2017-03-28, from http://www.journalism.org/2016/02/25/reddit-news-users-more-likely-to-be-male-young-and-digital-in-their-news-preferences/

Bero, L., & Rennie, D. (1995). The Cochrane Collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. *Jama, 274*(24), 1935–1938. Retrieved 2017-03-26, from http://jama.jamanetwork.com/article.aspx?articleid=393319

Bruckman, A., Curtis, P., Figallo, C., & Laurel, B. (1994). Approaches to managing deviant behavior in virtual communities. In *CHI Conference Companion* (pp. 183–184).

Campbell, D. T. (1981). Comment: Another perspective on a scholarly career. *Scientific inquiry and the social sciences*, 453–501.

Campbell, D. T. (1998). The experimenting society. In *The experimenting society: Essays in honor of Donald T. Campbell* (p. 35). New Brunswick: Transaction Publishers.

Campbell, D. T., & Stanley, J. C. (1963). *Experimental and Quasi-Experimental Designs for Research* (1st ed.). Wadsworth Publishing.

Casteel, H., Thakor, M., Johnson, R., & others. (2011). Human Trafficking and Technology: A framework for understanding the role of technology in the commercial sexual exploitation of children in the US. Retrieved 2017-03-28, from http://www.iu.edu/~traffick/_resources/_literature/_research/_assets/Human-Trafficking-and-Technology.pdf

Centivany, A., & Glushko, B. (2016). Popcorn Tastes Good: Participatory Policymaking and Reddit's. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 1126–1137). ACM. Retrieved 2017-03-29, from http://dl.acm.org/citation.cfm?id=2858516

Chalmers, I., Hedges, L. V., & Cooper, H. (2002, March). A Brief History of Research Synthesis. *Evaluation & the Health Professions, 25*(1), 12–37. Retrieved 2017-03-26, from http://dx.doi.org/10.1177/0163278702025001003 doi: 10.1177/0163278702025001003

Chan, M. (2017, March). ADL Tackles Hate Speech With Silicon Valley Command Center | Time.com. *Time Magazine*. Retrieved 2017-03-28, from http://time.com/4699823/adl-silicon-valley-hate-center/

Chancellor, S., Pater, J. A., Clear, T., Gilbert, E., & De Choudhury, M. (2016). # thyghgapp: Instagram Content Moderation and Lexical Variation in Pro-Eating Disorder Communities. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp. 1201–1213). ACM. Retrieved 2017-01-15, from http://dl.acm.org/citation.cfm?id=2819963

Cheng, J., Danescu-Niculescu-Mizil, C., & Leskovec, J. (2014, May). How Community Feedback Shapes User Behavior. *ICWSM 2014*. Retrieved 2017-03-29, from http://arxiv.org/abs/1405.1429 (arXiv: 1405.1429)

Citron, D., & Wittes, B. (2017, January). *Follow Buddies and Block Buddies: A Simple Proposal to Improve Civility, Control, and Privacy on Twitter.* Retrieved 2017-03-29, from https://lawfareblog.com/follow-buddies-and-block-buddies-simple-proposal-improve-civility-control-and-privacy-twitter

Citron, D. K., & Norton, H. L. (2011). Intermediaries and hate speech: Fostering digital citizenship for our information age. *Boston University Law Review, 91*, 1435. Retrieved 2017-01-15, from http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1764004

Contandriopoulos, D., Lemire, M., Denis, J.-L., & Tremblay, □. (2010, December). Knowledge Exchange Processes in Organizations and Policy Arenas: A Narrative Systematic Review of the Literature. *Milbank Quarterly, 88*(4), 444–483. Retrieved 2017-03-17, from http://onlinelibrary.wiley.com.libproxy.mit.edu/doi/10.1111/j.1468-0009.2010.00608.x/abstract doi: 10.1111/j.1468-0009.2010.00608.x

Cousins, J. B., & Whitmore, E. (1998). Framing participatory evaluation. *New directions for evaluation, 1998*(80), 5–23. Retrieved 2017-03-17, from http://onlinelibrary.wiley.com/doi/10.1002/ev.1114/full

Desposato, S. (2014). Ethical Challenges and Some Solutions for Field Experiments. Retrieved 2017-03-29, from http://desposato.org/ethicsfieldexperiments.pdf

Desposato, S. (2015). *Ethics and Experiments: Problems and Solutions for Social Scientists and Policy Professionals.* Routledge.

Dimond, J. P., Dye, M., Larose, D., & Bruckman, A. S. (2013). Hollaback!: the role of storytelling online in a social movement organization. In *Proceedings of the 2013 conference on Computer supported cooperative work* (pp. 477–490). ACM. Retrieved 2016-05-25, from http://dl.acm.org/citation.cfm?id=2441831

Doleac, J. L., & Stein, L. C. (2013). The visible hand: Race and online market outcomes. *The Economic Journal, 123*(572), F469–F492. Retrieved 2017-03-28, from http://onlinelibrary.wiley.com/doi/10.1111/ecoj.12082/full

Edelman, B. G., Luca, M., & Svirsky, D. (2016). Racial discrimination in the sharing economy: Evidence from a field experiment. Retrieved 2017-03-28, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2712393

Facebook. (n.d.). *Group Admin Basics: What is a Group Admin?* Retrieved 2017-03-29, from https://www.facebook.com/help/418065968237061/

Fiesler, C., Chancellor, S., Hoffmann, A. L., Pater, J., & Proferes, N. J. (2016). Challenges and Futures for Ethical Social Media Research. In *AAAI Conference on Web and Social Media (ICWSM): Workshop.*

Forte, A., Larco, V., & Bruckman, A. (2009). Decentralization in Wikipedia governance. *Journal of Management Information Systems, 26*(1), 49–72. Retrieved 2017-03-29, from http://www.tandfonline.com/doi/abs/10.2753/MIS0742-1222260103

Geiger, R. S. (2014). Successor Systems: The Role Of Reflexive Algorithms In Enacting Ideological Critique. *Selected Papers of Internet Research, 4.* Retrieved 2016-04-23, from http://spir.aoir.org/index.php/spir/article/view/942

Geiger, R. S. (2016, June). Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society, 19*(6), 787–803. Retrieved 2016-08-29, from http://dx.doi.org/10.1080/1369118X.2016.1153700 doi: 10.1080/1369118X.2016.1153700

Geiger, S. (2015). 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. Retrieved from https://cscwethics2015.files.wordpress.com/2015/02/geiger.pdf

Gerber, A. S., & Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation.* WW Norton.

Gillespie, T. (2010). The politics of 'platforms'. *New Media & Society, 12*(3), 347–364. Retrieved 2017-01-17, from http://nms.sagepub.com/content/12/3/347.short

Good, O. (2013, August). Does Your Gamertag Have Herpes? Beware Xbox Live Enforcement United. *Kotaku.* Retrieved 2015-09-23, from http://kotaku.com/does-your-gamertag-have-herpes-beware-xbox-live-enfor-1019141385

Grevet, C., & Gilbert, E. (2015). Piggyback Prototyping: Using Existing, Large-Scale Social Computing Systems to Prototype New Ones. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems* (pp. 4047–4056). ACM. Retrieved 2016-05-19, from http://dl.acm.org/citation.cfm?id=2702395

Grimmelmann, J. (2015). The law and ethics of experiments on social media users. Retrieved 2017-03-28, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2604168

Halfaker, A., Geiger, R. S., Morgan, J. T., & Riedl, J. (2012). The rise and decline of an open collaboration system: How Wikipedia's reaction to popularity is causing its decline. *American Behavioral Scientist*, 0002764212469365. Retrieved 2016-05-03, from http://abs.sagepub.com/content/early/2012/12/26/0002764212469365.abstract

Hill, B. M., & Shaw, A. (n.d.). Studying Populations of Online Communities. In *The Handbook of Networked Communication.* New York, NY: Oxford University Press. Retrieved 2017-03-25, from https://mako.cc/academic/hill_shaw-populations_of_communities-DRAFT.pdf

Hornbæk, K., Sander, S. S., Bargas-Avila, J. A., & Grue Simonsen, J. (2014). Is once enough?: on the extent and content of replications in human-computer interaction. In *Proceedings of the 32nd annual ACM conference on Human factors in computing systems* (pp. 3523–3532). ACM. Retrieved 2016-05-23, from http://dl.acm.org/citation.cfm?id=2557004

Irani, L., & Silberman, M. (2014). From critical design to critical infrastructure: Lessons from Turkopticon. *interactions*, *21*(4), 32–35. Retrieved 2016-05-25, from http://dl.acm.org/citation.cfm?id=2627392

Irani, L. C., & Silberman, M. (2013). Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 611–620). ACM. Retrieved 2016-05-25, from http://dl.acm.org/citation.cfm?id=2470742

Johnson, L. B. (1966, June). *296 - Memorandum on the Use and Management of Computers by Federal Agencies.* President of the United States. Retrieved 2017-01-25, from http://www.presidency.ucsb.edu/ws/index.php?pid=27677

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. *arXiv preprint arXiv:1511.06578*. Retrieved 2017-03-26, from https://arxiv.org/abs/1511.06578

Kiene, C., Monroy-Hernández, A., & Hill, B. M. (2016). Surviving an "Eternal September": How an Online Community Managed a Surge of Newcomers. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 1152–1156). New York, NY, USA: ACM. Retrieved 2017-03-29, from http://doi.acm.org/10.1145/2858036.2858356 doi: 10.1145/2858036.2858356

Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., & Pohlmann, N. (2013). Online Controlled Experiments at Large Scale. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 1168–1176). New York, NY, USA: ACM. Retrieved 2015-10-03, from http://doi.acm.org/10.1145/2487575.2488217 doi: 10.1145/2487575.2488217

Lomas, N. (2017, March). Twitter nixed 635k+ terrorism accounts between mid-2015 and end of 2016. *TechCrunch.* Retrieved 2017-03-28, from https://techcrunch.com/2017/03/21/twitter-nixed-635k-terrorism-accounts-between-mid-2015-and-end-of-2016/

MacKinnon, R. (2012). *Consent of the Networked: The Worldwide Struggle for Internet Freedom.* Basic Books.

Martinez, F. (2015, January 26). Cops Want Waze to Get Rid of Its Police-tracking Feature. *Fusion.* Retrieved 2017-03-28, from http://fusion.net/story/40459/one-cop-is-leading-a-crusade-to-ban-a-waze-feature-he-says-puts-police-in-danger/

Matias, J. N. (2016a, September). The Civic Labor of Online Moderators. Oxford, UK. Retrieved 2017-03-29, from http://ipp.oii.ox.ac.uk/sites/ipp/files/documents/JNM-The_Civic_Labor_of_Online_Moderators__Internet_Politics_Policy_.pdf

Matias, J. N. (2016b). Going Dark: Social Factors in Collective Action Against Platform Operators in the Reddit Blackout. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 1138–1151). ACM. Retrieved 2017-03-29, from http://dl.acm.org/citation.cfm?id=2858391

Matias, J. N. (2016c, October). *Posting Rules in Online Discussions Prevents Problems & Increases Participation.* Retrieved 2017-03-27, from http://civilservant.io/r_science_sticky_coments_1.html

Matias, J. N. (2016d, April). A toxic web: what the Victorians can teach us about online abuse. *The Guardian.* Retrieved 2017-03-26, from https://www.theguardian.com/technology/2016/apr/18/a-toxic-web-what-the-victorians-can-teach-us-about-online-abuse

Matias, J. N. (2017, February). *Persuading Algorithms With an AI Nudge.* Retrieved 2017-03-27, from https://civilservant.io/persuading_ais_preserving_liberties_r_worldnews.html

Matias, J. N., Johnson, A., Boesel, W. E., Keegan, B., Friedman, J., & DeTar, C. (2015). Reporting, Reviewing, and Responding to Harassment on Twitter.

arXiv preprint arXiv:1505.03359. Retrieved 2015-11-10, from http://arxiv.org/abs/1505.03359

Metz, R. (2017, March). Facebook Live's new suicide-prevention tools come with good intentions but many questions. *MIT Technology Review*. Retrieved 2017-03-28, from https://www.technologyreview.com/s/603772/big-questions-around-facebooks-suicide-prevention-tools/

Meyer, M. N. (2015, May). *Two Cheers for Corporate Experimentation: The A/B Illusion and the Virtues of Data-Driven Innovation* (SSRN Scholarly Paper No. ID 2605132). Rochester, NY: Social Science Research Network. Retrieved 2016-04-24, from http://papers.ssrn.com/abstract=2605132

Miguel, E., Camerer, C., Casey, K., Cohen, J., Esterling, K. M., Gerber, A., ... others (2014). Promoting transparency in social science research. *Science*, *343*(6166), 30–31. Retrieved 2017-03-25, from http://science.sciencemag.org/content/343/6166/30.short

Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., ... others (2015). Promoting an open research culture. *Science*, *348*(6242), 1422–1425. Retrieved 2017-03-25, from http://science.sciencemag.org/content/348/6242/1422.short

Oakley, A. (2000). Experiments in knowing: Gender and method in the social sciences. Retrieved 2016-12-17, from http://philpapers.org/rec/OAKEIK

O'Donovan, C. (2016, August). Nextdoor Rolls Out Product Fix It Hopes Will Stem Racial Profiling. *BuzzFeed*. Retrieved 2017-03-28, from https://www.buzzfeed.com/carolineodonovan/nextdoor-rolls-out-product-fix-it-hopes-will-stem-racial-pro

Popper, K. (1947). *The open society and its enemies*. Routledge.

Postigo, H. (2009, September). America Online volunteers. *International Journal of Cultural Studies*, *12*(5), 451–469. Retrieved 2015-08-19, from http://ics.sagepub.com.libproxy.mit.edu/content/12/5/451

Reisman, D. (2016, May). *A Peek at A/B Testing in the Wild*. Retrieved 2016-05-30, from https://freedom-to-tinker.com/blog/dreisman/a-peek-at-ab-testing-in-the-wild/

Rheingold, H. (1993). *The virtual community: Homesteading on the electronic frontier.* MIT press.

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry.* Retrieved 2017-03-29, from https://pdfs.semanticscholar.org/b722/7cbd34766655dea10d0437ab10df3a127396.pdf

Seltzer, W. (2010). Free speech unmoored in copyright's safe harbor: Chilling effects of the DMCA on the first amendment. *Harv. JL & Tech., 24*, 171. Retrieved 2017-03-28, from http://heinonline.org/hol-cgi-bin/get_pdf.cgi?handle=hein.journals/hjlt24&section=7

Sunstein, C. R. (2009). *Republic. com 2.0.* Princeton University Press.

Tan, Z. Y. (2016, August). Hospitals Are Partnering With Uber to Get Patients to Checkups. *The Atlantic.* Retrieved 2017-03-28, from https://www.theatlantic.com/health/archive/2016/08/hospitals-are-partnering-with-uber-to-get-people-to-checkups/495476/?utm_source=atlfb

Thakor, M., & others. (2013). Networked trafficking: reflections on technology and the anti-trafficking movement. *Dialectical Anthropology, 37*(2), 277–290. Retrieved 2017-03-28, from http://link.springer.com/article/10.1007/s10624-012-9286-6

Thakor, M. N. (2016). *Algorithmic detectives against child trafficking: data, entrapment, and the new global policing network* (Doctoral dissertation, Massachusetts Institute of Technology). Retrieved 2017-03-28, from https://dspace.mit.edu/handle/1721.1/107039

Vitak, J., Shilton, K., & Ashktorab, Z. (2016). Beyond the belmont principles: Ethical challenges, practices, and beliefs in the online data research community. In *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work & Social Computing* (pp. 941–953). ACM. Retrieved 2017-03-29, from http://dl.acm.org/citation.cfm?id=2820078

Wark, L. (2016, November). Inside Alphabet's Jigsaw, the powerful tech incubator that could reshape geopolitics. *Quartz.* Retrieved 2017-03-28, from

https://qz.com/846836/inside-google-jigsaw-the-powerful-tech-incubator-that-wants-to-reshape-geopolitics/

Weiss, C. H. (1977). Research for policy's sake: The enlightenment function of social research. *Policy analysis*, 531–545. Retrieved 2017-03-17, from http://www.jstor.org/stable/42783234

Weiss, C. H. (1979). The many meanings of research utilization. *Public administration review*, *39*(5), 426–431. Retrieved 2017-03-17, from http://www.jstor.org/stable/3109916

Zhang, A. X., Verou, L., & Karger, D. (2017). Wikum: Bridging Discussion Forums and Wikis Using Recursive Summarization. Retrieved 2017-04-19, from http://people.csail.mit.edu/axz/papers/wikum.pdf

Chapter5

Preventing Online Harassment with Community-Led Policy Experiments

In this chapter, I report results from a 14-million subscriber science discussion community that tested the effects of posting the rules at the top of discussions. In an experiment during August and September 2016, we randomly assigned rule messages to half of the community's 2,214 discussions and question-answer sessions. I found that posting the rules increased newcomer rule compliance by over 7 percentage points on average in the community, from 75% to 82%.

When Stephen Hawking faced insults and jokes about his medical condition and private life in a 2015 Q&A on reddit, the abusive comments violated policies in r/science, a community of over 13 million subscribers that was hosting the conversation. Many readers "downvoted" the comments to make the less prominent, and the community's volunteer moderators also noticed. After moderators determined that the comments violated community policies against insults, abusive language, and jokes, they quickly removed the comments from the conversation. A year later in July 2016, moderators removed over 130,000 comments or posts, banned 460 people from the community, and made 558 adjustments to the community's policy documents in one month.

Because volunteer moderators create and enact policy for millions of people in data-rich online environments, they have a unique opportunity to conduct policy evaluations independently of the platforms that host their communities. In this study, moderators and I attempted to prevent online harassment by prominently displaying community policies in discussions among the the 13 million subscribers of r/science. Our intervention increased norm-compliance among first time commenters by 7.4 percentage points and also increased participation rates by 30.1%. Overall, the intervention can prevent thousands of incidents of harassment each month. Our experiment provides an example of the potential of online moderators to form an "experimenting society" of community-led policy evaluation in the governance of digitally-mediated behavior.

Policy Evaluation in The Experimenting Society

In the early 1970s, as US policymakers were conducting the first federal randomized trials of government policy, the evaluation methodologist Donald Campbell imagined a democracy where policy experiments could be a common form of civic participation. In this "experimenting society," local communities would evaluate their policies systematically. Imagining communities as 'co-agents directing their own society," Campbell suggested that they participate in the analysis and decisionmaking associated with policy evaluation. Communities could deliberate on interventions, debate dependent variables, and question analysis procedures. Campbell also argued that groups could improve the validity of policy knowledge through experimental disputation: replicating, cross-validating, and debating ideas for reducing social ills through experimental means

(Campbell, 1998). Yet while participatory field experiments have occasionally been attempted in public health research (Ammerman et al., 2003; Pazoki, Nabipour, Seyednezami, & Imami, 2007; Mishra, Luce, & Baquet, 2009), the idea of community-led policy evaluation has mostly remained a thought experiment (Oakley, 2000; John, Smith, & Stoker, 2009).

Online platforms have created conditions amenable to the idea of an experimenting society by collecting large-scale data on human behavior, delegating the governance of that behavior to community volunteers, and sharing the data back to communities. Large-scale online data collection has already opened new avenues of enquiry in the social sciences (Lazer et al., 2009), and platform operators like Microsoft routinely conduct as many as 300 randomized trials per day (Kohavi, Longbotham, Sommerfield, & Henne, 2009). Most of these experiments are applied to pricing and advertising, and their results are usually retained as trade secrets by platforms. In some cases, revelations of this corporate social experimentation has been greeted with public fear and outrage (Grimmelmann, 2015a).

Platforms are not the only ones who possess the building blocks for conducting policy experiments online. On many platforms, moderators are given access to substantial data about the behavior in their communities, usually to support communities to manage semi-automated moderation software (Geiger & Ribes, 2010; Geiger & Halfaker, 2013). Using these application programming interfaces (APIs), communities can deploy software that monitors behavior and coordinates social responses to that behavior (Geiger, 2014, 2016). These APIs can also be used to deploy community-led field experiments.

Community Policymaking Online

For over 40 years, volunteer moderators have created and enacted policies within online communities that range from a dozen people to tens of millions (Butler, Sproull, Kiesler, & Kraut, 2002; Grimmelmann, 2015b). In 1970s Berkeley, librarians and record shop staff managed the *Community Memory* system that hosted local community discussions and classified ads (Bruckman, 1998). In the 1980s, *conference hosts* at the WELL, BBS *SysOps*, and Usenet *moderators* discussed rules, maintained order, and mediated among community participants (Bruckman, Curtis, Figallo, & Laurel, 1994; Rheingold, 1993). In the 1990s, as online conversation became a thriving business model, AOL organized tens

of thousands of volunteer *community leaders* to manage its chatrooms (Postigo, 2003). Currently, these volunteer governance roles are played by *administrators* on Wikipedia (*Wikipedia:Administrators*, 2015), group *admins* (Facebook, n.d.) on Facebook, *moderators* on Slashdot (Lampe & Resnick, 2004), *group organizers* on Meetup (Lai, 2014), *enforcement united* on Xbox (Good, 2013), and *subreddit moderators* on the social news platform reddit (Matias, 2016a).

On reddit, over fifty thousand "subreddit" communities of up to 15 million subscribers are moderated by volunteer teams that can grow to over a thousand people per team. Communities include job boards, news discussion groups, markets, mental health support communities, breaking news, and book clubs (Massanari, 2015; Leavitt & Clark, 2014). Each community on reddit hosts group discussions about links, media, and original content that community members share. As other community members respond and vote on submissions, the platform's algorithms identify popular material across the site and promote it to a series of algorithmically-generated pages that a crowdsourced "front page of the internet." Within communities, moderators are given the role because they founded the community, are appointed by other moderators, or are selected through community-led elections and recruitment process (Matias, 2016a).

To learn about cultures of governance on reddit, I have spent over a year and a half as a digital ethnographer on the platform (Boellstorff, Nardi, & Pearce, 2012). I observed and participated in hundreds of communities, analyzed large-scale patterns in behavioral data on the site, and used those findings to guide interviews, conversations, and participation as a moderator in several communities. My findings revealed that communities on reddit already have a flourishing culture of data analysis and policy debate. Community policies on reddit tend to focus on acceptable behavior and content. For example, the r/science community that hosted professor Hawking in 2015 prohibits "abusive comments," medical advice, personal anecdotes, and jokes. They also require that all submissions for discussion link to peer-reviewed publications. Actions that violate these policies are removed by moderators, and "repeat or flagrant offenders" can be banned by moderators from the community.

Beyond responsive measures such as removing comments, volunteer moderators also have substantial powers to shape the design and infrastructure of their communities to prevent problems. Moderators can define the visual appearance of the reading experience, adjust the function of the platform's voting

system, and modify automated bots that observe and intervene in conversations. For example, r/science has a policy requiring people to label the discipline of any link that is shared with the community. When a new link is shared, an automated bot reviews the link, offers feedback to the contributor if the submission does not comply with community policies, and removes it if the post is not amended properly. Moderators of r/science and other subreddits collect systematic data about their work and publish transparency reports that invite wider community debate about their decisions (Matias, 2016a).

Moderators routinely share policy ideas and moderation practices across communities. Many communities share common, open source software to coordinate their efforts within a community. Moderators participate in shared groups to debate policies and best practices. Policies also spread between communities that share common moderators. In June 2015, moderators of r/science held 331 moderation positions in other communities. That summer, moderators revealed the extent of their capacity to coordinate when over 2,200 communities joined a strike that successfully forced the reddit company to expand support for its volunteer moderators (Matias, 2016b).

Conversations with moderators from 2015-2016 revealed a common need for communities to systematically test their policies. In the summer of 2016, I designed the CivilServant software, which coordinates interventions and collects data for community-led randomized trials of moderation policies. This experiment with r/science on policies designed to prevent online abuse is the first of what I hope will grow into an experimenting policy culture on reddit and other platforms. Over the coming years, I hope to support thousands of new field studies and community replications on behavioral policy in digitally-mediated environments.

How Can We Increase Newcomer Rule Compliance while Preserving Their Participation Rates?

Many harassing and abusive comments in the r/science community come from people who are participating for the first time. In July 2016, moderators removed 494 newcomer comments per day, 39.1% of all the comments they removed on average and 52.3% of all newcomer comments. First-time comments were also more likely to violate community policies than contributions from experienced moderators. Theories of newcomer socialization in social psychol-

ogy would expect that first-time participants are in a "discovery phase" where they are still learning the norms of a community and deciding if they want to participate further (Levine, Moreland, & Choi, 2001). They may not yet be aware of community policies against abusive language, insulting jokes, and personal anecdotes in discussions of scientific publications.

In this study, moderators hoped to improve newcomer behavior by posting the rules to the top of discussions (hypothesis 1). Rule postings influence what Robert Cialdini has called *injunctive norms*, people's awareness of rules that "specify what ought to be done... through the promise of social sanctions" (Cialdini, Kallgren, & Reno, 1991). By influencing people's awareness of community norms, moderators hoped to increase the chance that newcomer comments follow the rules. Field experiments elsewhere have found that increasing the visibility of rules has affected littering behavior (Reiter & Samuel, 1980; De Kort, McCalley, & Midden, 2008), smoking in hotel rooms (Dawley, Morrison, & Carrol, 1981), environmental conservation by hotel guests (Goldstein, Cialdini, & Griskevicius, 2008), and crime reporting (Bickman & Green, 1977).

Moderators and I also wondered if increasing norm compliance could reduce community growth (hypothesis 2), which moderators wished to avoid. While people aren't likely to change hotels or park elsewhere after reading a sign about environmental policies, it's much easier to leave an online community. Theories from computer science and group identity predict that posting the rules may also reduce participation in the community overall (Figure 5-1). If the rules convince a newcomer that their comment isn't acceptable, they may never comment at all. Posted rules also increase the complexity of the task of commenting, requiring newcomers to tailor their comment to the community. Since people are less likely to complete tasks that are more complex, we expected that fewer newcomers will comment when the rules are visible (Eickhoff & de Vries, 2011). Furthermore, since newcomers encounter rules during the investigation phase of their relationship with a community, they might decide that they don't fit in this community and may never participate (Levine et al., 2001). On the other hand, one lab experiment found that people expressed a greater intent to participate in online conversations that they considered to be well-moderated. Since prior research led us to coflicting expectations, we decided to monitor newcomer participation rates alongside rule compliance.

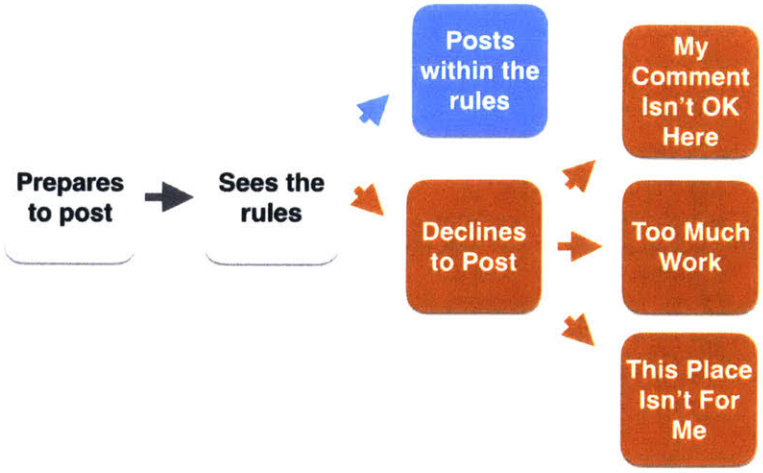


Figure 5-1: While posting rules may increase newcomer compliance with social norms, it may also reduce the overall number of contributions from newcomers

How I Designed Policy Experiments Together With An Online Community

Within the r/science community, moderators wanted to evaluate the effect of showing the rules on newcomer norm compliance and participation rates. As a researcher, I wanted to learn how online communities make sense of policy evaluations that they direct. As I describe the methods of this experiment, I also tell the story of the conversations with community members that shaped its design.

Moderators of r/science first contacted me about working together after I hosted a discussion on reddit in February 2016 about testing community policies. A week later, we held our first discussion about policies to test, using real-time chat software with moderators in different geographies. When one moderator said, "I want to know the impact of a sticky comment explaining the rules," the group quickly agreed. Over the next hour, moderators discussed experiment procedures and debated possible outcome variables—as researchers, they were experienced at thinking about measurements and modeling.

The intervention moderators agreed on was a "sticky comment" that moderators sometimes use to pin a list of rules to the top of a discussion. During the experiment, the CivilServant software new discussions as they were posted and determined whether or not they were a Q&A with a prominent scientist (like the discussion with Stephen Hawking). Based on the type of discussion,

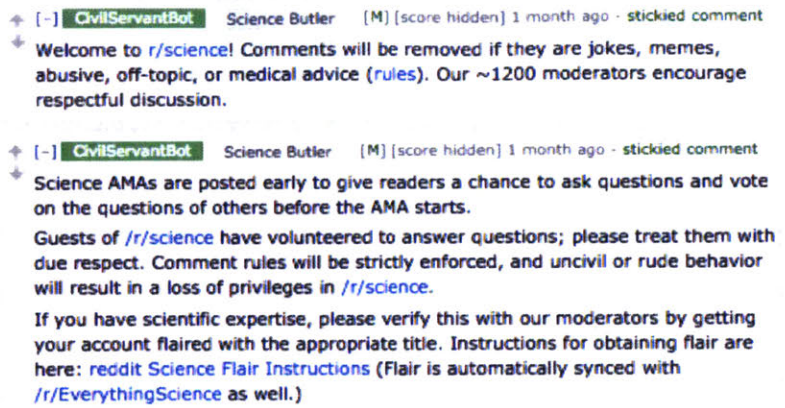


Figure 5-2: In the experiment, the CivilServant software automatically posted these sticky comments to the top of different discussions

the software used block randomization to either post the rules to a discussion or withhold posting the rules. Moderators adjusted the design of the community so that they would not be able to see which discussions included sticky comments, blinding themselves to the intervention.[1]

I developed the text for the sticky comment together with moderators by posting a set of suggestions for the kinds of information that the sticky comments could include: descriptions and links to policies, information about the consequences of violating policies, information about the number of moderators, and a welcome message. Using a collaborative text editor, moderators developed a series of candidate messages before collectively choosing the final versions (Figure 5-2).

In community conversations, we realized that observing the effect of sticky comments on newcomer rule-compliance would require three merged datasets: the history of community participation, the history of moderation actions, and new community activity as it occurred.

Together, we defined *newcomers* as accounts who have not previously contributed in the community in the last six months before the study began. To generate this value, we used an aggregate list of all accounts that contributed public comments to the subreddit over a six month period.[2] Since moderators were blinded to the sticky comments, we automatically removed all replies to

[1] A pre-analysis plan is available at `https://osf.io/knb48/files/osfstorage/57bef819594d9001fcd0e193/`

[2] Details on the construction of this variable at `https://github.com/c4fcm/CivilServant-Analysis/blob/master/FrontPageRScienceDataCreation.ipynb`

the rule postings and omitted those replies from the dependent variables.

To evaluate the effect on norm compliance, the software observed *the moderation outcome for a newcomer comment.* In the lifecycle of a single comment, it may be removed by moderators or an automated moderation bot. Removed comments can also be restored by moderators, sometimes after appeal by the person who made the comment. This study considered the final state of the comment after the experiment concluded. The large number of moderators in the r/science community maintain continuous monitoring of comments that appear in their community and often discuss difficult cases. The moderation team also enforces consistent moderator compliance with policies. For example, during the experiment, one moderator was discovered to have exceeded the community policies. Their position was stripped from them and the history of their past actions was reviewed and corrected by other moderators. The unit of observation for this outcome is a comment by a newcomer made on a post that falls within the experiment sample. Not all discussions included newcomer comments. Among all of the 2,214 treated discussions, 1,872 discussions included at least one comment and 830 included at least one newcomer comment.

To evaluate the effect on participation rates, the software observed *the number of newcomers per discussion.* The unit of observation for this outcome is a *top-level post* to the community, and the measure is the number of newcomers who make comments, omitting replies to the sticky comments.

Estimating the Outcomes

When we discussed the best way to estimate the effect of sticky comments on newcomers' rule compliance and participation rates, moderators were skeptical about the comparability of different discussions. Based on years of experience, moderators knew that the audience and behavior of commenters varied widely by the time of day, day of week, topic, and the relative popularity of a post. They pointed out that when a discussion is promoted by reddit's popularity algorithms, it often reaches a much wider group of people, many of whom do not understand the group's norms. While randomization handles this variation in theory, I also set the software to observe several regression adjustment variables to improve the precision of the final estimates (Gerber & Green, 2012). Before starting the experiment, I observed these variables over a period of sev-

eral weeks to decide which modeling approaches and adjustment variables were appropriate for each question. [3]

Many discussion submissions are rejected by moderators, at which point they are no longer open to further commenting. To adjust for moderator removals, we observed the final state of *post visibility* at the end of the experiment. Some discussions reach the very top of reddit's rankings for a community due to their popularity. These posts often achieve even wider readerships. To monitor the relative popularity of posts, the software sampled the rankings of r/science every five minutes. *Minutes in top five* is the log-transformed number of minutes that the post appeared in the top five items in the subreddit rankings. During the experiment, 21.9% of discussions featured in the top five, with many remaining at the top for days. Some discussions featured *live Q&A* sessions with a notable personality. These discussions often reach the widest readerships, and moderators sometimes offer greater attention to these "Ask-Me-Anything" (AMA) discussions. The experiment included 24 Q&As. The software also observed the *post hour* and whether the discussion was started on a *weekend.*

To model the effect of sticky comments on rule compliance, I used a random intercepts logistic regression that adjusted for post visibility, whether it was a live Q&A, and the number of minutes the discussion spent in the top five of the community's popularity rankings. The random intercepts model allows the model to adjust for any differences in the moderation of discussions that varied in these ways; it also provides an accurate estimate of the effect on individual comments when our intervention applied to whole discussions (Singer & Willett, 2003). To model the effect of sticky comments on the commenting rate of newcomers, I used a zero-inflated poisson regression that predicts the incidence rate of comments. Zero-inflation accounts for the high number of discussions that received no newcomer comments (Long & Freese, 2014). For example, a discussion that is removed quickly by moderators is very unlikely to receive any comments. The final model uses a post's visibility and its time in the rankings to predict cases of no newcomer comments. The rest of the model adjusts for post visibility, whether it was a live Q&A, 23 different discussion topics, whether the discussion was started on the weekend, and the

[3] full details of preliminary modeling results are available at `https://rawgit.com/mitmedialab/CivilServant-Analysis/master/reports/experiment.planning.07.16.2016.html`

hour that the discussion was opened.

The Effects of Posting Rules to the Top of Discussions

I evaluated the effect of sticky comments in r/science from Aug 25, 2016 to Sep 23, 2016. After removing 24 observations in randomization blocks that were spoiled by software errors, the final results included a total of 24 question-answer discussions and 2,190 discussions of academic publications.[4] The 20,385 newcomer comments were 29.7% of all comments in this period.

Without posting the rules, a first-time commenter has a 75.2% chance of complying with community norms. I found that posting the rules has a positive 7.3 percentage point effect on the chance that a newcomer's comment will be allowed to remain by moderators on average in r/science, holding all else constant (Table 5.1, Figure 5-3). This finding is consistent with prior research on the effect of increasing the visibility of social norms.

In the secondary analysis, I found that rather than reducing participation, posting the rules increases the incidence rate of newcomer comments by 38.1% on average, holding all else constant (Table 5.2, Model 1). Yet this effect is not consistent across all conversations. Since the system block-randomized between Q&A sessions and other discussions, I could test for different effects in different kinds of conversations. In this followup analysis, I discovered that while the rule postings increased the rate of comments in discussions of published research by 59%, they caused a 65.5% reduction in the rate of comments on question-answer discussions with researchers (Table 5.2, Model 2).

Why do we see the opposite effect between Q&A sessions and more ordinary discussion threads? Some prior research led us to expect decreases in participation rates and others led us to expect increases. Those theoretical questions are left open when we observe opposite outcomes in the field. Community members and I have developed several potentially-testable explanations that future research should consider:

Differences in outcomes might be explained by differences in the messages used for Q&A discussions and more common conversations (see figure 5-2). The longer Q&A message also asks participants to share information about

[4]Because data collection was taking weeks longer than expected, I implemented the stop rule, even though the effect size was not as high as the 20 percentage point effect size cutoff specified in the pre-analysis plan.

	Null Model	Model 1	Final Model
Treatment			0.44**
			(0.16)
Post Visible		1.06***	1.09***
		(0.19)	(0.18)
Live Q&A		0.74*	0.79*
(AMA)		(0.36)	(0.36)
ln Top 5		−0.12***	−0.12***
Minutes		(0.02)	(0.02)
(Intercept)	0.29***	0.24	0.02
	(0.08)	(0.14)	(0.16)
AIC	23775.05	23723.28	23717.64
BIC	23790.90	23762.90	23765.18
Deviance	22477.45	22480.89	22489.44
Log Likelihood	-11885.53	-11856.64	-11852.82
Num. obs.	20385	20385	20385
Num. groups:			
discussion	830	830	830
Variance: discussion			
(Intercept)	2.72	2.48	2.40
Variance: Residual	1.00	1.00	1.00

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05$

Table 5.1: Posting rules increased the chance that newcomer comments would comply with the rules by 7.3 percentage points (Final Model). The effect applied differently to different kinds of discussions. Results of a random-intercepts logistic regression predicting the visibility of comments from an experiment in r/science. This dataset includes 830 discussions and 20,385 comments (Figure 5-3).

their academic expertise. It's possible that this emphasis on credentials may have dissuaded people from participating.

The effect of posting rules may also operate differently at different scales. During the community debriefing, one commenter posed a theory to explain why posting rules had caused such a large increase in participation. Explaining that certain parts of the reddit platform show how many comments are associated with a discussion, they pointed out that the intervention increments the count: "I bet that the rules comment increases participation because it makes it say (1 comment) on the forum index so people click the link to read the comment." Q&A discussions are widely promoted and receive many comments; this effect would not apply to them.

This study has several limitations. First, it's possible that the effects I observed come from changes in behavior. Yet since the software cannot observe non-commenters, it's also possible that the intervention influenced who par-

	Null Model	Model 1	Model 2
Zero Model	0.83***	−2.68***	−2.54***
(Intercept)	(0.07)	(0.39)	(0.38)
Zero Mode:	0.37***	1.01**	0.95**
Post Visible	(0.10)	(0.33)	(0.32)
Zero Model:	−0.23***	0.07*	0.06*
ln Top 5 Mins	(0.01)	(0.03)	(0.03)
Treatment		0.32***	0.46***
		(0.01)	(0.02)
Treatment x			−1.53***
Live Q&A			(0.06)
Post Visible		−0.57***	−0.58***
		(0.02)	(0.02)
Live Q&A		0.26***	0.83***
(AMA)		(0.03)	(0.03)
(22 Omitted)			
(Topics)			
ln Top 5 Mins		0.52***	0.52***
		(0.00)	(0.01)
Weekend Post		0.23***	0.26***
		(0.02)	(0.02)
Post Hour		0.22***	0.22***
		(0.01)	(0.01)
Post Hour ^{2}		−0.01***	−0.01***
		(0.00)	(0.00)
(Intercept)	3.21***	−2.76***	−2.84***
	(0.01)	(0.14)	(0.14)
AIC	88772.86	40465.64	39722.94
Log Likelihood	-44382.43	-20199.82	-19827.47
Num. obs.	2214	2214	2214

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05$

Table 5.2: Posting rules increased the incidence rate of newcomer comments by 38.1% on average (Model 1). But the effect applied differently to different kinds of discussions. Rule postings increased the rate of comments in discussions of published research by 59% and caused a 65.5% reduction in the rate of comments on question-answer discussions with researchers (Model 2). This table shares results of a zero-inflated poisson regression predicting the incidence rate of comments in 2,214 discussions from an experiment in r/science.

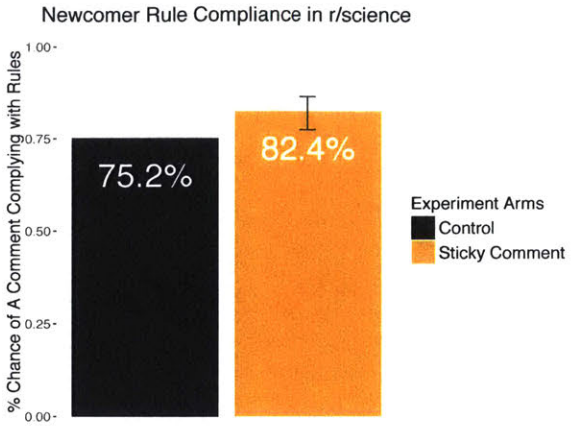


Figure 5-3: Posting rules increased newcomer rule compliance by 7.3 percentage points on average (n = 2,214 randomizations, 20,385 newcomer comments) (Table 5.1).

ticipated, not how they participated. Second, although the models for the first hypothesis assume independence, comments are not independent from each other; many comments are replies to other comments. Third, the study randomized at the level of discussions rather than individual commenters, which results in greater uncertainty in the model (see error bars on Figure 5-3). Finally, some of the regression adjustment variables are correlated with each other, which may also slightly influence the results. Many of these limitations are the results of the novel approach I take to conduct policy evaluations outside control of the platform software. To make claims that are squarely within the valid implications of the study design, I have described this policy intervention as harassment prevention rather than behavior change, which would require individual-level interventions rather than discussion-level ones.

Community Debriefing on Policy Implications and Study Ethics

When I shared the outcome of the experiment with the community in October 2016, the day-long conversation received 478 comments and 14,354 votes. The post was read over 200,000 times and was ranked one of the most popular discussions on the reddit platform for over a day. Community members discussed policy decisions, suggested new ways to design interventions, shared personal experiences, debated the experiment design, and asked questions about experiment ethics.

When debating the policy implications of the study, community members discussed the goals that might lead them to adopt or reject showing the rules in the future. One commenter asked, "what if lack of conflict & increased participation is bad?" Another wondered, "can this cause censorship if taken to an extreme?"[5] Participants debated whether norms about acceptable speech should be called censorship at all.

When discussing the design of the intervention, one person imagined that "the wording is extremely important." Sharing personal anecdotes about communities where they had avoided making comments, this person contrasted the welcoming wording in the study with harsher language that they thought might deter newcomer participation. People who shared personal stories tended to be outliers who wanted their experience to be part of the policy evaluation. One person remarked, "I don't think I've ever read any community rules ever."

Finally, community members asked about research ethics. Some asked about informed consent and requested details of the MIT ethics review of the study design. One person argued strongly against the study, in a discussion that revealed deeper disagreements with moderators over the commenting rules. Describing the community's policies against abusive language "censorship," this person argued that moderators should encourage people to speak in any way they liked. Since this person saw the idea of community policies as unethical, they also held the experiment to be unethical, since it extended the influence of those policies.[6]

Policy Impact Among reddit Communities

At the time of writing, the r/science community is planning to start applying sticky comments with rules to all conversations in the community. As conflict grew in the r/politics discussion group of 3 million subscribers, they began posting the rules to the top of discussion groups. At the time of writing, three more communities are planning community replications of findings in r/science. While this first experiment required substantial software development to observe variables important to communities and coordinate interventions, the software is now general enough for a new community to deploy a replication study in a single afternoon.

[5]Quotations have been obfuscated to protect research participants. In research on public, searchable online platforms, complete quotes make participants easily identifiable.

[6]I received no requests to opt out of this study.

Discussion

Back in 2015, if moderators had known to make commenting policies more visible at the top of the discussion, Stephen Hawking may have been greeted with fewer cruel and harassing jokes. As the group best positioned to understand and respond to social problems in their community, volunteer moderators were able to quickly remove the offending comments. Together with the CivilServant project, moderators were also able to develop and evaluate policy interventions that can reduce the rates of harassment received by Stephen Hawking and others.

In the r/science community, posting rules to the top of discussions could prevent 1,838 people from engaging in online harassment and other unacceptable behavior each month on average (Table 5.1, Final Model). While the effect on newcomer participation rates vary between different kinds of conversations, moderators would still gain 9,631 new participants each month on average (Table 5.2, Model 1).

In "The Experimenting Society," Donald Campbell imagined a future where communities could shape design and analysis of their own policy experiments, as well as participate in wider policy debates through replication and re-analysis. The findings offer a practical, working example of a community-led policy experiment in the governance of online behavior. From the design of the variables to discussions over the meaning of the results, participants in r/science shaped and debated our policy experiment. They are also the ones who will ultimately decide if they want to implement the intervention that I tested. In time, other communities may replicate and extend these early findings. As digital communications systems continue to simplify data collection and the deployment of large-scale policy interventions governing social behavior, I hope that community-led evidence-based policy can become more common.

References

Ammerman, A., Corbie-Smith, G., St. George, D. M. M., Washington, C., Weathers, B., & Jackson-Christian, B. (2003). Research expectations among African American church leaders in the PRAISE! project: a randomized trial guided by community-based participatory research. *American Journal of Pub-*

lic Health, 93(10), 1720–1727. Retrieved 2017-02-05, from http://ajph.aphapublications.org/doi/abs/10.2105/AJPH.93.10.1720

Bickman, L., & Green, S. K. (1977, March). Situational Cues and Crime Reporting: Do Signs Make a Difference?1. *Journal of Applied Social Psychology, 7*, 1–18. Retrieved 2016-03-18, from http://onlinelibrary.wiley.com.libproxy.mit.edu/doi/10.1111/j.1559-1816.1977.tb02413.x/abstract doi: 10.1111/j.1559-1816.1977.tb02413.x

Boellstorff, T., Nardi, B., & Pearce, C. (2012). *Ethnography and virtual worlds: A handbook of method.* Princeton University Press.

Bruckman, A. (1998). Finding one's own in cyberspace. *High Wired: On the Design, Use, and Theory of Educational MOOs. Ed. Cynthia Haynes and Jan Rune Holmevik. Ann Arbor, MI: U of Michigan P*, 15–24. Retrieved 2015-09-23, from http://cumincad.architexturez.net/system/files/pdf/59c3.content.pdf

Bruckman, A., Curtis, P., Figallo, C., & Laurel, B. (1994). Approaches to managing deviant behavior in virtual communities. In *CHI Conference Companion* (pp. 183–184).

Butler, B., Sproull, L., Kiesler, S., & Kraut, R. (2002). Community effort in online groups: Who does the work and why. *Leadership at a distance: Research in technologically supported work*, 171–194.

Campbell, D. T. (1998). The experimenting society. In *The experimenting society: Essays in honor of Donald T. Campbell* (p. 35). New Brunswick: Transaction Publishers.

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). 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.

Dawley, H. H., Morrison, J., & Carrol, S. (1981, January). The Effect of Differently Worded No-Smoking Signs on Smoking Behavior. *International Journal of the Addictions, 16*(8), 1467–1471. Retrieved 2016-03-18, from http://dx.doi.org/10.3109/10826088109039197 doi: 10.3109/10826088109039197

De Kort, Y. A., McCalley, L. T., & Midden, C. J. (2008). Persuasive trash cans: Activation of littering norms by design. *Environment and Behavior*. Retrieved 2016-03-17, from http://eab.sagepub.com/content/early/2008/05/11/0013916507311035.short

Eickhoff, C., & de Vries, A. (2011). How crowdsourcable is your task. In *Proceedings of the workshop on crowdsourcing for search and data mining (CSDM) at the fourth ACM international conference on web search and data mining (WSDM)* (pp. 11–14). Retrieved 2016-06-28, from http://www.academia.edu/download/30680905/csdm2011_proceedings.pdf#page=11

Facebook. (n.d.). *Group Admin Basics: What is a Group Admin?* Retrieved 2015-09-23, from https://www.facebook.com/help/418065968237061/

Geiger, R. S. (2014). Bots, bespoke, code and the materiality of software platforms. *Information, Communication & Society, 17*(3), 342–356. Retrieved 2017-02-05, from http://www.tandfonline.com/doi/abs/10.1080/1369118X.2013.873069

Geiger, R. S. (2016, June). Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society, 19*(6), 787–803. Retrieved 2016-08-29, from http://dx.doi.org/10.1080/1369118X.2016.1153700 doi: 10.1080/1369118X.2016.1153700

Geiger, R. S., & Halfaker, A. (2013). When the levee breaks: without bots, what happens to Wikipedia's quality control processes? In *Proceedings of the 9th International Symposium on Open Collaboration* (p. 6). ACM. Retrieved 2017-02-05, from http://dl.acm.org/citation.cfm?id=2491061

Geiger, R. S., & Ribes, D. (2010). The work of sustaining order in wikipedia: the banning of a vandal. In *Proceedings of the 2010 ACM conference on Computer supported cooperative work* (pp. 117–126). ACM. Retrieved 2016-08-28, from http://dl.acm.org/citation.cfm?id=1718941

Gerber, A. S., & Green, D. P. (2012). *Field experiments: Design, analysis, and interpretation.* WW Norton.

Goldstein, N. J., Cialdini, R. B., & Griskevicius, V. (2008, October). A Room with a Viewpoint: Using Social Norms to Motivate Environmental Conservation in Hotels. *Journal of Consumer Research, 35*(3), 472–482. Retrieved 2016-03-18, from http://jcr.oxfordjournals.org/content/35/3/472 doi: 10.1086/586910

Good, O. (2013, August). Does Your Gamertag Have Herpes? Beware Xbox Live Enforcement United. *Kotaku.* Retrieved 2015-09-23, from http://kotaku.com/does-your-gamertag-have-herpes-beware-xbox-live-enfor-1019141385

Grimmelmann, J. (2015a). The Law and Ethics of Experiments on Social Media Users. Retrieved 2017-01-15, from http://papers.ssrn.com/sol3/Papers.cfm?abstract_id=2604168

Grimmelmann, J. (2015b, April). *The Virtues of Moderation* (SSRN Scholarly Paper No. ID 2588493). Rochester, NY: Social Science Research Network. Retrieved 2015-06-24, from http://papers.ssrn.com/abstract=2588493

John, P., Smith, G., & Stoker, G. (2009). Nudge nudge, think think: two strategies for changing civic behaviour. *The Political Quarterly, 80*(3), 361–370. Retrieved 2017-01-15, from http://onlinelibrary.wiley.com/doi/10.1111/j.1467-923X.2009.02001.x/full

Kohavi, R., Longbotham, R., Sommerfield, D., & Henne, R. M. (2009). Controlled experiments on the web: survey and practical guide. *Data mining and knowledge discovery, 18*(1), 140–181. Retrieved 2016-01-20, from http://link.springer.com/article/10.1007/s10618-008-0114-1

Lai, C.-H. (2014). Can our group survive? An investigation of the evolution of mixed-mode groups. *Journal of Computer-Mediated Communication, 19*(4), 839–854. Retrieved 2017-02-05, from http://onlinelibrary.wiley.com/doi/10.1111/jcc4.12075/full

Lampe, C., & Resnick, P. (2004). Slash (dot) and burn: distributed moderation in a large online conversation space. In *Proceedings of the SIGCHI conference on Human factors in computing systems* (pp. 543–550). ACM. Retrieved 2015-06-24, from http://dl.acm.org/citation.cfm?id=985761

Lazer, D., Pentland, A. S., Adamic, L., Aral, S., Barabasi, A. L., Brewer, D., ... others (2009). Life in the network: the coming age of computational social science. *Science (New York, NY), 323*(5915), 721. Retrieved 2016-04-24, from http://www.ncbi.nlm.nih.gov/pmc/articles/pmc2745217/

Leavitt, A., & Clark, J. A. (2014). Upvoting Hurricane Sandy: Event-based News Production Processes on a Social News Site. In *Proceedings of the 32nd Annual ACM Conference on Human Factors in Computing Systems* (pp. 1495–1504). New York, NY, USA: ACM. Retrieved 2015-09-25, from http://doi.acm.org/10.1145/2556288.2557140 doi: 10.1145/2556288.2557140

Levine, J. M., Moreland, R. L., & Choi, H.-S. (2001). Group socialization and newcomer innovation. *Blackwell handbook of social psychology: Group processes, 3*, 86–106.

Long, J. S., & Freese, J. (2014). *Regression Models for Categorical Dependent Variables Using Stata, Third Edition* (3edition ed.). College Station, TX: Stata Press.

Massanari, A. L. (2015). *Participatory Culture, Community, and Play: Learning from Reddit* (No. 75). Peter Lang Publishing Inc.

Matias, J. N. (2016a, September). The Civic Labor of Online Moderators. Oxford, UK. Retrieved from http://ipp.oii.ox.ac.uk/sites/ipp/files/documents/JNM-The_Civic_Labor_of_Online_Moderators__Internet_Politics_Policy_.pdf

Matias, J. N. (2016b). Going Dark: Social Factors in Collective Action Against Platform Operators in the Reddit Blackout. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 1138–1151). New York, NY, USA: ACM. Retrieved 2016-08-31, from http://doi.acm.org/10.1145/2858036.2858391 doi: 10.1145/2858036.2858391

Mishra, S. I., Luce, P. H., & Baquet, C. R. (2009). Increasing Pap smear utilization among Samoan Women: Results from a community based participatory randomized trial. *Journal of health care for the poor and underserved, 20*(2 Suppl), 85. Retrieved 2017-02-05, from https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3604739/

Oakley, A. (2000). Experiments in knowing: Gender and method in the social sciences. Retrieved 2016-12-17, from http://philpapers.org/rec/OAKEIK

Pazoki, R., Nabipour, I., Seyednezami, N., & Imami, S. R. (2007). Effects of a community-based healthy heart program on increasing healthy women's physical activity: a randomized controlled trial guided by Community-based Participatory Research (CBPR). *BMC Public Health, 7*, 216. Retrieved 2017-02-05, from http://dx.doi.org/10.1186/1471-2458-7-216 doi: 10.1186/1471-2458-7-216

Postigo, H. (2003). Emerging sources of labor on the Internet: The case of America Online volunteers. *International review of social History, 48*(S11), 205–223. Retrieved 2015-09-23, from http://journals.cambridge.org/abstract_S0020859003001329

Reiter, S. M., & Samuel, W. (1980). Littering as a Function of Prior Litter and The Presence or Absence of Prohibitive Signs. *Journal of Applied Social Psychology, 10*, 45–55. Retrieved 2016-03-18, from http://onlinelibrary.wiley.com/doi/10.1111/j.1559-1816.1980.tb00692.x/abstract

Rheingold, H. (1993). *The virtual community: Homesteading on the electronic frontier.* MIT press.

Singer, J. D., & Willett, J. B. (2003). *Applied longitudinal data analysis: Modeling change and event occurrence.* Oxford university press.

Wikipedia:Administrators. (2015, September). Retrieved 2015-09-23, from https://en.wikipedia.org/w/index.php?title=Wikipedia:Administrators&oldid=681136693 (Page Version ID: 681136693)

Chapter6

AI Nudges: Reducing the Algorithmic Promotion of Unreliable News by Influencing Social Behavior

In this chapter, I report results from an experiment led by reddit's world news discussion group, which had 16 million subscribers in December 2016. Moderators in this community were concerned about the interactions between human behavior and reddit algorithms that spread misleading and sensationalized tabloid news. They wished to intervene in ways that pro-socially influenced human and algorithm behavior while preserving individual liberties.

In this experiment, possibly the first systematic effort to evaluate the effects of human nudges on machine behavior, we encouraged commenters to fact-check unreliable news or fact-check and vote on the articles. Compared to no action at all, I found that both interventions increased the chance that individual commenters would link to further evidence in discussions. Surprisingly, I found that while encouraging fact-checking could reduce the promotion of unreliable news by reddit's popularity algorithms, I failed to find an effect from encouraging fact-checking and voting. As black box algorithms and AI systems play a greater role in human affairs, similar experiments may help us govern the unexpected interactions between human and machine behavior.

Introduction

Well-functioning democracies require widespread citizen awareness of reliable news (Lippmann, 1946; Gans, 2003; Schudson, 2000). Even as communications technologies have broadened citizen access to misinformation, artificial intelligence systems now exercise substantial influence over the news that citizens read and share (Pariser, 2011). Consequently, the work of maintaining democratic societies now involves managing algorithms that influence the popularity of misinformation (Sunstein, 2009). As algorithm operators and their systems make decisions about which information to remove or promote, they face pressure to protect society from the risks of misinformation while also preserving individual liberties from the risks of censorship (Gillespie, 2010; MacKinnon, 2012).

Contemporary information technologies such as news aggregators may increase the visibility of unreliable but popular news. Aggregators observe activity and ratings from a population and present ranked lists of "trending" suggestions (Resnick & Varian, 1997). These aggregators can enter into feedback loops with social behavior when an aggregator increases human actions that influence the aggregator's ranking algorithms. For example, when people perceive low-quality cultural products as popular, the social influences related to those beliefs can "lead the herd astray," making those products more popular over time even if the popularity information is initially false (Salganik & Watts, 2008). These feedback loops of rapidly-building social and algorithmic attention have been linked with escalating patterns of online conflict and political turbulence (Massanari, 2015; Margetts, John, Hale, & Yasseri, 2015).

Because the workings of recommender systems are typically secret (Diakopoulos, 2016), interventions that respond to unreliable news may unknowingly create feedback loops that increase rather than reduce the spread of misinformation. As an alternative to censoring misinformation, online communities could address misinformation by commenting with factual corrections in discussions of unreliable news, since citizens tend to update their beliefs when presented with factual corrections of misleading news headlines (Wood & Porter, 2016). As a behavioral nudge, fact-checking might dampen the effects of misinformation while preserving individual liberties (Sunstein & Thaler, 2003). Yet behavioral nudges might also cause an "AI nudge" that might not be beneficial. Fact-checking might increase the recommendation ranking of unreliable news,

spreading misinformation further if an aggregator system interprets increased activity from fact-checking as evidence of greater popularity.

With algorithmic accountability, online communities might be able to anticipate algorithmic side effects. Algorithmic transparency might require online platforms to publicly-document details from the software code of aggregators and their training data (Diakopoulos, 2016). If that were not possible, researchers might conduct algorithmic audits, closely controlling the inputs to a system to observe patterns in its outputs (Sandvig, Hamilton, Karahalios, & Langbort, 2014). Yet since human feedback loops with a recommender system constitute the system's full outcome, even perfect knowledge of an algorithm is insufficient for predicting the full effect of a social intervention on cumulative human and machine behaviors.

The societal and political risks from algorithmically-promoted unreliable news illustrate the importance of estimating the second-order effects of collective human behavior on news aggregators online. Here I report the results of an "AI Nudge," a large-scale field experiment estimating the effect of encouraging fact-checking on human and machine promotion of unreliable news online. First, I find that encouraging fact-checking increased fact-checking behavior and reduced the ranking of unreliable news by as many as 4 positions over time on average, in a large news discussion community on the reddit platform. Yet encouraging community participants to fact-check and influence the aggregator did not cause a discernable effect on rank position over time. Second, I show that the effect on the aggregator of encouraging fact-checking follows a cubic polynomial curve over time. Finally, I discuss the AI nudge as a novel method for preserving individual liberties while managing risks from macro patterns of algorithm-societal interaction.

Methods

To the best of my knowledge, this is the first field experiment estimating the effect of pro-socially influencing human social behavior on the related behavior of a black box system whose design is unknown to the experimenters. To conduct this experiment, I collaborated with moderators of the r/worldnews community on the reddit platform, a group that shares, comments, and votes on

the relative quality of news articles about places other than the United States.[1] When the experiment began, this English language community had over 14 million subscribers. In a six-week period from mid-September 2016 into October, 914 articles per day were submitted to the community on average, 2.4% of which were from unreliable news sites that community members tended to report for being more sensational and less reliably sourced. Of all of these articles, 46% were permitted by moderators. Even articles removed my moderators were viewed and discussed by community participants, receiving a median of 2 comments and a mean of 20, since several hours can elapse before moderators decide to remove an article.

Experiment Procedure

I conducted the field experiment with CivilServant, a software agent that accessed the reddit platform through the reddit application programming interface (API). Moderators granted CivilServant moderator-level access, allowing it to collect near real-time information on community activity and automatically-post the experiment interventions. The software queried reddit for information about news articles every sixty seconds and information about their ranking position every four minutes. Articles were included in the experiment if the news link was from a website domain that moderators considered a frequently-unreliable source.[2]

To conduct this experiment, a software program identified recently-submitted news articles from sources considered to be unreliable by volunteer moderators of a world news discussion community with over 14 million subscribers on the reddit platform. The software randomly assigned these articles to receive one of three conditions: no action, a persistent message encouraging readers to fact-check the article by commenting with links to alternative evidence (Figure 6-1), and a persistent message that encouraged readers to fact-check and consider down-voting the article to reduce its position in the rankings (Figure 6-2).

In the control condition the software took no action. In the treatment conditions, the software posted persistent messages to the top of conversations.

[1] An experiment pre-analysis plan is available at `https://osf.io/knb48/files/osfstorage/583b0a37594d9000441f6d76/`

[2] selected domains included dailymail.co.uk, express.co.uk, mirror.co.uk, news.com.au, nypost.com, thesun.co.uk, dailystar.co.uk, metro.co.uk

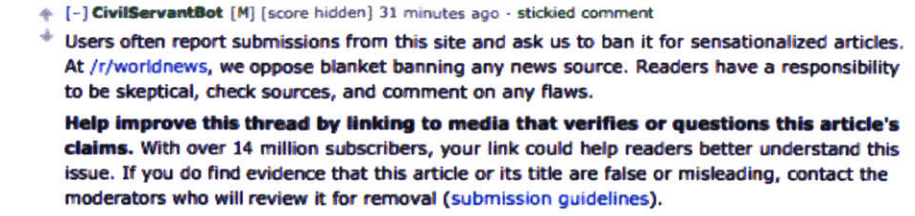


Figure 6-1: Treatment A: Encouraging Fact-Checking Behavior. This message was one of two possible messages posted to the top of discussions of news articles from unreliable news articles submitted to the reddit community.

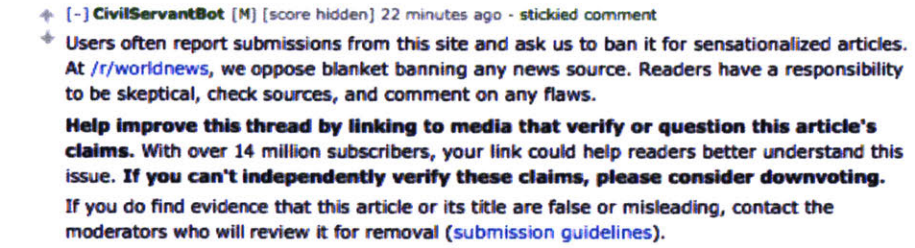


Figure 6-2: Treatment B: Encouraging Fact-Checking and Voting Behavior. This message was identical to the message in Figure 6-1, with an added encouragement to vote on articles that readers could not verify.

These messages would always be displayed as the top-most comment to anyone reading the discussion. In the fact-checking condition, the message asked commenters to share links to alternative evidence about the story being discussed (Figure 6-1). In the fact-checking and voting condition, the message included an added sentence encouraging readers to down-vote low-quality news articles (Figure 6-2). Experiment conditions were block-randomized by time into balanced groups of 12, applying each arm four times in each block. Blocks where software errors prevented full application of the treatment or observation of variables were removed.

Data Collection

The /r/worldnews community and CivilServant software began the study on November 27, 2016. Since the reddit platform changed the behavior of their ranking algorithms partway into the study,[3] this analysis includes 1,104 unreliable news posts from December 7, 2016 to February 15, 2017. During this time, the system made observations of news articles, comments in discussions

[3] see `https://www.reddit.com/r/announcements/comments/5gvd6b`

	Control	Fact-Checking (A)	Fact-Checking and Voting (B)	Total
Articles with comments	291	300	278	869
Articles permitted	47%	44%	42%	44%
Comments	22,286	4,550	8,254	35,090
Comments with links	805	249	405	1,459

Table 6.1: Characteristics of data collected about comments in discussions of unreliable news

threads responding to those articles, and the rank position of those articles over time.

Participants made 35,090 experiment-eligible comments in 869 news discussions among the 1104 discussion threads that received an arm of the experiment (Table 6.1). In the analysis, I omit 345 comments made by five common automated systems. I also omit any comment made as part of the experiment interventions. When comments linked to other reddit discussions or to media from image-hosting domains that rarely publish factual information, these 2,773 comments were labeled as not including links to further information. Within eligible comments, the system observed a binary value of whether a comment included at least one link to further evidence. A typical comment with evidence would respond directly to the intervention with a list of links to news articles, occasionally with explanatory text. These comments occasionally attracted further discussion of which of these links were trustworthy.

For each news article submitted, the software observed the time that the discussion was opened. The software also observed when any article was removed or reinstated by moderators at any time through the end of the experiment period. In post and comment level analyses, the system recorded a binary variable for the final visibility of an article discussion at the conclusion of the experiment. For analyses of news article rankings at a time, the system recorded the visibility state of the article discussion at the observed time (Figure 6-3).

To observe the second-order effect of encouraging fact-checking on the behavior of the reddit news recommendations, the system took samples every four minutes of the top 100 recommendations made by reddit on the world-news community's default aggregator, called the "hot" ranking on reddit. The system also expanded the ranking observations to include the top 300 recommendations for 516 articles from January 13, 2017 to the end of the experi-

	Observed Top 100 Ranked Articles			Observed Top 300 Ranked Articles		
Start Date	Dec 7, 2016			Jan 13, 2017		
End Date	Feb 15, 2017			Feb 15, 2017		
Article Count	1104			516		
	Min	Mean	Max	Min	Mean	Max
Median score of other items	12	42.1	175.5	13.5	52.7	175.5
Sample Timing Offset (Minutes)	-0.5	-0.004	0.5	-0.5	0.002	0.5
Treatment articles in top N (A)	0	0.6	4	0	1.7	6
Treatment articles in top N (B)	0	0.7	4	0	2	7

Table 6.2: Summary statistics for ranking snapshots used to estimate the effects of encouraging fact-checking and voting on the rank position of a news article

ment. I used these observations to create two longitudinal datasets of the rank position of every news link for the 24 hour period after it was first submitted (Table 6.2). The software observed the ranking position of each news link over 7 hours, 105 times. In these two datasets, the measure of rank position ranges from 0 to 100 in one and from 0 to 300 in the other. A value of zero indicates that the news article did not appear in the top N during that observation. A value of 1 indicates the least prominent rank position, and a value of 100 or 300 indicates most prominent position (Figure 6-3). Because data collection expanded to include 300 ranking items partway through the experiment, data on rankings in the top 300 represents a subset of randomization blocks within the larger sample of articles.

Within ranking data, the system also observed several variables used for regression adjustment. Since rankings positions are relative to other articles, the data includes adjustment variables for the median reddit score of other articles in the top N subreddit rankings at that observation (M_{Nt}). The ranking data also includes information on small timing offsets in ranking observations that adjust for variation in software sampling operations (TO_t). The system also observed two variables used to adjust for spillover effects: the number of other news articles in the top N that received the first treatment (OA_{Nt}) and the number of other news articles in the top N that received the second treatment (OB_{Nt}) (Table 6.2).

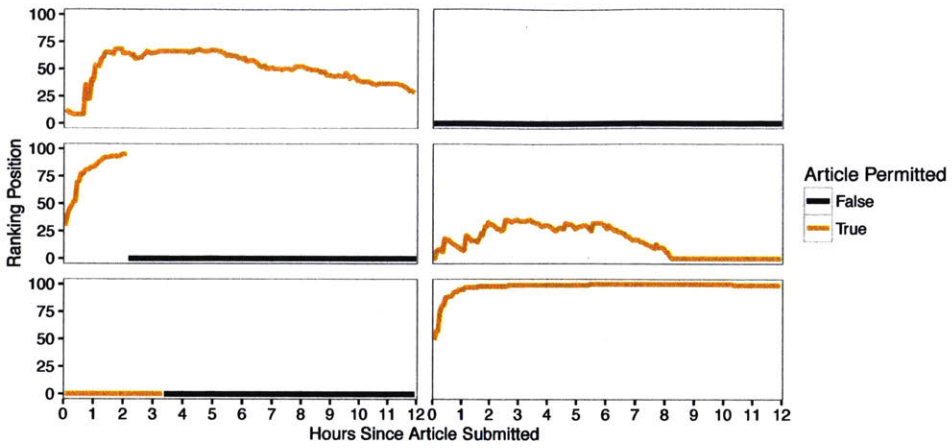


Figure 6-3: Rank position over time for six example unreliable news articles. A rank position of 0 indicates that the news article was not featured in the 100 articles recommended by reddit's algorithms within the subreddit. Observed ranks range from 1 to 100, with 100 as the most prominent. When an article is removed by moderators, it is removed from the rankings and receives a value of 0 for the remaining time.

Results

In a series of logistic regression models, I found as expected that both encouragements to fact-check increased the chance that individual comments would include links to evidence in discussions of unreliable news sources (Figure 6-4 (A)). Both interventions also increased the chance that individual discussions would have at least one comment that included links to further evidence (Figure 6-4 (B)). Since the encouragements towards fact-checking successfully influenced commenters' responses to unreliable news articles, I was able to observe the second-order effects of this social influence on the behavior of the news aggregation algorithm.

To observe the effect of both interventions on the news aggregation algorithm, I observed an article's rank position every four minutes in the top 100 and 300 items suggested by the default aggregator algorithm in this community. I then fit a series of linear regression models, one for each four-minute period in the first seven hours after an article was posted. This method allows me to model the average treatment effect without making assumptions about the shape of the curve taken by an article through the rankings, using an approach employed by Taylor, Muchnik, and Aral in randomized trials with recommender systems on Facebook (Taylor, Muchnik, & Aral, 2014). I ex-

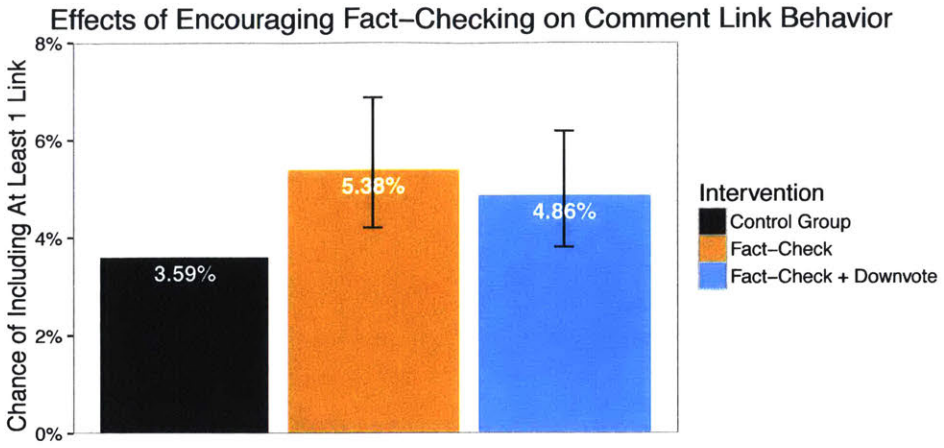


Figure 6-4: In a series of logistic regression models, encouraging fact-checking increased the chance that individual comments would include links (n= 35090 comments in 869 discussions, p = 0.0014, Table 6.3)

pected that encouraging fact-checking might increase the ranking of unreliable news articles and that encouraging voting alongside fact-checking would make dampen that effect. To the contrary, encouraging fact-checking caused unreliable news articles to be ranked as many as 4 positions lower in the top 100 items than control group articles, at the height of the effect. In contrast, I fail to find an effect from encouraging voting with fact-checking. Both results are consistent with findings among the top 300 recommended articles (Figure 6-5).

Because news articles begin at a similar position in aggregator rankings, rise in some cases to prominence, and subsequently recede, the effects on rankings at a moment of time can be modeled with a cubic polynomial curve over the log-transformed age of the discussion (Table 6.6). In the early minutes after a news article is submitted, the average treatment effect is small enough to be unobservable in this sample. The effect on rank position grows in the first forty-five minutes, quickly reaches a maximum effect size and declines towards zero over time (Figure 6-6).

Analysis

I estimated the average treatment effect on the chance of a comment including links to evidence with a logistic regression of comments, adjusting standard errors using the maximum-likelihood Huber-White method for comments clustered within discussions that received the treatment (Figure 6-4, Table 6.3)

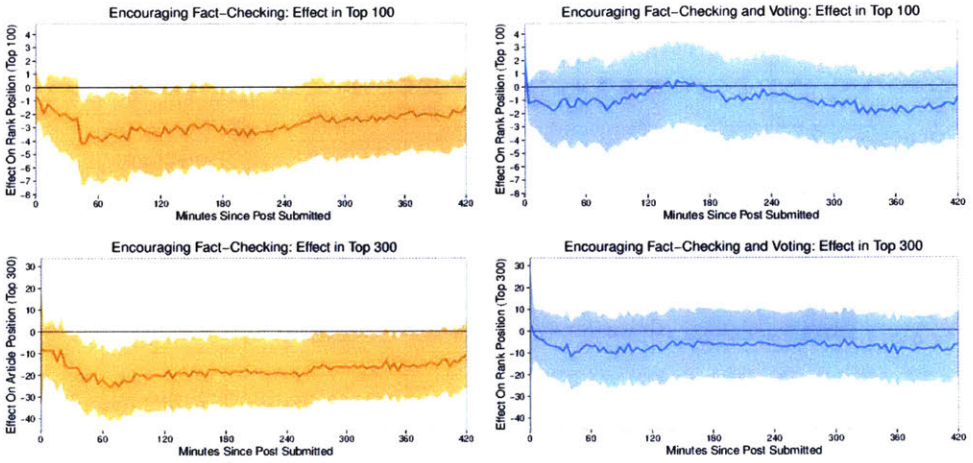


Figure 6-5: While encouraging fact-checking reduces the rank position of an unreliable news article, adding an encouragement to vote has no discernable effect on rank position at any time. This chart shows the effect size and 95% confidence intervals of 105 linear regression models estimating the average treatment effect on rank position at a moment in time after a post was submitted.

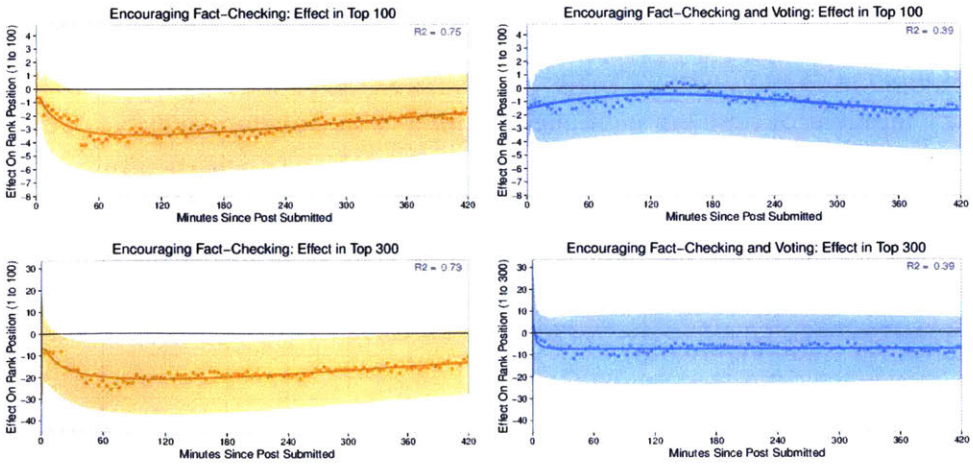


Figure 6-6: The effect of encouraging fact-checking on relative ranking positions over time followed a cubic polynomial curve. This chart shows the fitted effect sizes and fitted 95% confidence intervals predicted by cubic polynomial models (Table 6.6).

(White, 1982; Huber, 1967). In a secondary analysis, I estimated the average treatment effect on the chance of a discussion to include at least one link-bearing comment with a logistic regression model (Table 6.4). In a linear regression model predicting differences in the maximum rank achieved by a news article, I failed to reject the null hypothesis of a difference between treatment and control groups on average (Table 6.5). This null result is likely due to the inability of a non-longitudinal model to adjust for the non-independence of individual articles and interference.

To estimate the effect of the interventions on the relative aggregator ranking of a news article at a period in time, I fit two sets of 105 linear regression models, one model each four-minute sample in the first seven hours after an article was submitted. The models follow the form $Position_{Nt} = \alpha + \beta_1 TA_t + \beta_2 TB_t + \beta_3 P_t + \beta_4 M_{Nt} + \beta_5 TO_t + \beta_6 OA_{Nt} + \beta_7 OB_{Nt} + \epsilon$ where N represents the number of observation within the range of ranking items observed: 100 or 300.

	Base Model	Comment Model
Encourage Fact-Checking		0.423**
		(0.132)
Encourage Fact-Checking + Voting		0.314*
		(0.130)
Intercept	−3.027***	−3.257***
	(0.071)	(0.103)
Article Permitted	−0.151	−0.032
	(0.121)	(0.100)
Num. obs.	35090	35090
Pseudo R^2	0.001	0.005
L.R.	6.261	46.679

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05$

Table 6.3: Encouraging fact-checking at the top of an online discussion of tabloid news increased the chance that a comment would include links to further evidence. Standard errors in this logistic regression are adjusted using the maximum-likelihood Huber-White method for comments clustered within discussions that received the treatment.

	Discussion Model
Encourage Fact-Checking	0.458**
	(0.170)
Encourage Fact-Checking + Voting	0.352*
	(0.172)
Intercept	−1.285***
	(0.141)
Permitted	−0.027
	(0.138)
AIC	1277.689
BIC	1297.716
Log Likelihood	-634.845
Deviance	1269.689
Num. obs.	1104

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05$

Table 6.4: Encouraging fact-checking at the top of an online discussion of unreliable news increased the chance that at least one comment would include links to further evidence.

Discussion

Democratic societies increasingly need effective measures to manage the spread of unreliable news by humans and artificial intelligence systems while also preserving individual liberties. Social interventions such as fact-checking avoid censorship, but these interventions can cause unanticipated second-order effects through artificial intelligence systems whose behaviors are opaque to the public. In this study, I demonstrate the second-order effects of a social intervention on the behavior of a news recommendation aggregator of unknown design. If the effects of other "AI Nudges" can also be observed in the field, they may offer a productive avenue for managing the societal risks from other feedback loops in human and machine behavior.

Contrary to the initial hypotheses, encouraging fact-checking caused unreliable news to receive lower rankings from the aggregator. Adding encouragement to down-vote those articles caused any effect to be indistinguishable from zero. Because this study observes the second-degree effects of encouraging certain social behaviors on the behavior of an algorithmic system, the explanation for this outcome likely results from some combination of human

	Base Model	Main Model
Intercept	32.96***	35.08***
	(3.05)	(3.28)
Article Permitted	−6.25***	−6.20***
	(1.85)	(1.85)
Hour Posted	−0.29	−0.28
	(0.54)	(0.54)
Hour Posted ^{2}	0.00	0.00
	(0.02)	(0.02)
Weekend	2.21	2.28
	(2.16)	(2.16)
Encourage Fact-Checking		−3.97
		(2.25)
Encourage Fact-Checking + Voting		−2.93
		(2.25)
R^2	0.01	0.02
Adj. R^2	0.01	0.01
Num. obs.	1104	1104
RMSE	30.52	30.50

$^{***}p < 0.001$, $^{**}p < 0.01$, $^{*}p < 0.05$

Table 6.5: Linear regression model predicting the maximum 24 hour rank of unreliable news articles fails to observe an effect from encouraging fact-checking or fact-checking and voting. This failure may result from the inability of this model to account for regression adjustment variables of the relative ranking of other items and overspill from other treatment posts in the rankings.

	Fact-Check Effect (100)	Fact-Check Effect (300)
Intercept	-0.724^{*}	-7.342^{***}
	(0.353)	(1.759)
ln Minutes	1.245^{***}	6.127^{***}
	(0.357)	(1.782)
ln Minutes^2	-0.921^{***}	-4.516^{***}
	(0.116)	(0.579)
ln Minutes^3	0.113^{***}	0.552^{***}
	(0.011)	(0.055)
R^2	0.745	0.733
Adj. R^2	0.738	0.725
Num. obs.	106	106
RMSE	0.362	1.804

$^{***}p < 0.001, ^{**}p < 0.01, ^{*}p < 0.05$

Table 6.6: Cubic polynomial models of the average treatment effect over time of encouraging fact-checking on position of a news article in the top 100 and top 300 ranked items. The shape of these effects and 95% confidence intervals are illustrated in Figure 6-6.

and machine behavior. Among humans, this outcome may arise from psychological reactance, a resistance to suggestions from authority (Brehm & Brehm, 2013). Alternatively, if news submitters worry that their links might receive negative votes, they might ask others to promote the article to balance out the voting behavior of readers. Without access to voting records held by the reddit platform, neither these theories of human voting behavior nor theories of machine responses to voting can be tested systematically.

References

Brehm, S. S., & Brehm, J. W. (2013). *Psychological reactance: A theory of freedom and control.* Academic Press.

Diakopoulos, N. (2016). Accountability in algorithmic decision making. *Communications of the ACM, 59*(2), 56–62.

Gans, H. J. (2003). *Democracy and the News.* Oxford University Press.

Gillespie, T. (2010). The politics of 'platforms'. *New Media & Society, 12*(3), 347–364.

Huber, P. J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 221–233).

Lippmann, W. (1946). *Public opinion* (Vol. 1). Transaction Publishers.

MacKinnon, R. (2012). *Consent of the Networked: The Worldwide Struggle for Internet Freedom.* Basic Books.

Margetts, H., John, P., Hale, S., & Yasseri, T. (2015). *Political turbulence: how social media shape collective action.* Princeton University Press.

Massanari, A. (2015). # Gamergate and The Fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media & Society.*

Pariser, E. (2011). *The filter bubble: How the new personalized web is changing what we read and how we think.* Penguin.

Resnick, P., & Varian, H. R. (1997). Recommender systems. *Communications of the ACM, 40*(3), 56–58.

Salganik, M. J., & Watts, D. J. (2008). Leading the herd astray: An experimental study of self-fulfilling prophecies in an artificial cultural market. *Social psychology quarterly, 71*(4), 338–355.

Sandvig, C., Hamilton, K., Karahalios, K., & Langbort, C. (2014). Auditing algorithms: Research methods for detecting discrimination on internet platforms. *Data and discrimination: converting critical concerns into productive inquiry.*

Schudson, M. (2000). *The good citizen: A history of American civic life.* JSTOR.

Sunstein, C. R. (2009). *Republic. com 2.0.* Princeton University Press.

Sunstein, C. R., & Thaler, R. H. (2003). Libertarian paternalism is not an oxymoron. *The University of Chicago Law Review,* 1159–1202.

Taylor, S. J., Muchnik, L., & Aral, S. (2014). Identity and opinion: A randomized experiment. Retrieved 2017-05-14, from `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2538130`

White, H. (1982). Maximum likelihood estimation of misspecified models. *Econometrica: Journal of the Econometric Society*, 1–25.

Wood, T., & Porter, E. (2016, August). *The Elusive Backfire Effect: Mass Attitudes' Steadfast Factual Adherence* (SSRN Scholarly Paper No. ID 2819073). Rochester, NY: Social Science Research Network.

Chapter7

The Uses of Community Experiments in Online Policy

Community-led experiment results can only serve an open society if they are distributed, debated, and used. In this chapter, I report the outcomes of adding experimental knowledge to the network of policy discussions and practices on reddit. I observe forces that shaped the ten-year history of a single policy adopted by reddit's politics community during the 2016 election. Drawing lessons from the field of policy research utilization, I also follow the spread and uses of evidence from two community-led experiments that I conducted with CivilServant.

Research evidence on reddit follows many of the patterns observed in government policymaking, yet the unique characteristics of platforms enable experiments to achieve rapid, widespread use in community policy. Research informs the personal practices of individual moderators, contributes to community deliberations, prompts new community research, and can influence platform designers to test platform-wide systems. I conclude this chapter with lessons from CivilServant for a society that uses experimental knowledge in community policymaking.

Appeals for civility during U.S. political discussions in 2016 must have seemed hopelessly naive. Candidates at the highest level of U.S. elections regularly adopted extreme language to attack groups of people and their political enemies, leading group chants that encouraged antagonism and violence. Throughout the year, many pointed to online communities as key nodes in feedback networks of polarization that fostered new forms of antagonism and grew from the conflicts that they amplified.

In September 2016, when moderators of reddit's politics discussion group began posting automated notices calling for "civil discussion" among their three million subscribers, political conversations on the platform had gained a reputation for extreme conflict. On one hand, large groups were organizing to generate and spread racist, misogynyst narratives more widely (Singal, 2016; Williams, 2016; Phillips, Beyer, & Coleman, 2017). On the other hand, a greater proportion of reddit users supported candidates who opposed this perspective (Barthel, Stocking, Holcomb, & Mitchell, 2016; Gottfried, Barthel, & Mitchell, 2017). Political conversations on reddit were ripe for the kinds of fierce, unstable chain reactions of participation described by Margetts and others in a recent book about political turbulence online (Margetts, John, Hale, & Yasseri, 2015), fuelled by conflicts to influence the site's news recommendation algorithms (Menegus, 2016).

How could the thirty-three volunteers who facilitated one of the largest battlegrounds on the English-language internet believe that asking people to be civil would make any difference? Automated civility messages were just one of the ideas that moderators tried as they managed a rapidly-growing, argumentative crowd of 138,000 monthly commenters who made roughly 1.8 million comments in September. Yet the automated messages attracted strong debate each time they appeared— which was every single discussion. "Why do you feel it's necessary? It's annoying, and trolls are going to be trolls," complained one person in the next month's open discussion with moderators.

The politics discussion group's moderators defended their policy of posting civility reminders by pointing to "great stats and research" from a field experiment conducted by another reddit community. These findings matched what moderators saw in their daily efforts to prevent the worst excesses of a volatile U.S. political conversation. Moderators reported that "we've noticed a measurable reduction in incivility." One moderator wrote that "you would be surprised how many people are *shocked* that their comment violated the rules,"

claiming that civility reminders "legitimately educate new users." Convinced partly by research and partly through their own experiences, moderators continued to post civility messages throughout the election, even as personal attacks and conflict continued at high volumes. Article links in the largest politics discussions attracted over a hundred thousand conflicting upvotes and downvotes from readers. Even with civility reminders, moderators would regularly remove over a thousand comments from discussions, up to ten percent of everything said.

Whatever its contribution to U.S. politics, this decision by moderators was a landmark in the relationship between internet users and behavioral research. For perhaps the first time in the forty-five year history of online moderation, community leaders justified a decision by citing a policy experiment designed and conducted by communities themselves. As the researcher who supported that study, I was encouraged by their decision, which validated my belief that transparent, community-led experiments can be a valuable resource for community governance online. I developed novel software to test that idea, and successfully conducted two large community experiments in 2016.

To my surprise, the first users of this evidence weren't the communities who conducted the studies. Both communities continued for months without deciding the implications of their own research. Instead, the politics community's debates over civility reminders provided the first evidence that the findings would be useful to communities. The ability to conduct community-led experiments could not alone ensure that communities would develop their own evidence-based policies. Communities would also need to incorporate the evidence into collective decisions about how to govern themselves.

In this chapter, I ask how ideas about moderation policies are spread, discussed, adopted, and rejected across the network of communities on the reddit platform. To explore this question, I follow the history of innovation and arguments over civility reminders in communities. I then look at how knowledge of our experiment was spread and used by communities. I ask these questions through interviews, content analysis, and data analysis of moderation practices in thousands communities across the platform. I find that while communities on reddit use experimental knowledge in similar ways to governments, communities are also connected by software architectures that allow a policy idea to spread quickly through deliberation and through code. I conclude with lessons for fostering widespread, policy evaluation practices that influence debates and

decisions about community governance online.

Civility Reminders: Policy Evolution Across Communities

To study the ways that policy ideas develop among online communities, I used a wide range of participatory and ethnographic methods between June 2015 and April 2017. I spent regular time in reddit communities, participated in months of internal conversations with moderators of seven communities, talked with over twenty moderators in semi-structured interviews, and analyzed discussions of moderator decisions in hundreds of communities. I also read public, historical records that revealed the development of policies within communities. In 2015, I observed a large-scale strike by moderators who were pressuring the company to expand their governance tools, interviewing over a dozen moderators from communities on both sides of the strike. In 2016, I worked with nine communities to design novel field experiments to answer governance questions of importance to them. At the time of writing, two of these studies were complete.

When the moderators of reddit's politics discussion decided to post automated civility reminders during the 2016 election season, they were building on a ten-year history to manage behavior on the platform. The designers of reddit created the first "subreddit" community in 2006, a pornography and violence oriented group for "NSFW" material. After creating new communities for language groups over the next two years, the company began supporting volunteers to create and moderate their own communities in 2008. Moderators were expected to recruit participants, develop their own norms, and decide for themselves how to enforce those norms. Moderators were given wide control over their communities, including the visual display of the interface, permitted contributors, and content removal. Late in 2009, the platform invited moderators to display guidelines in a "sidebar" on the right side of community pages (*Pictures and Images*, 2009).

By delegating power to volunteer moderators, reddit was continuing a system of community governance with origins in the early internet. In 1970s Berkeley, librarians and record store staff managed terminals from the Community Memory system, which also charged participants 25 cents in hopes of deterring abusive behavior (Bruckman, Curtis, Figallo, & Laurel, 1994). In the 1980s, people used modems and their university networks to access community-

governed social spaces. Across the 80s social internet, WELL *conference hosts*, BBS *SysOps*, and UseNet *moderators* created and enacted local policies in thousands of communities (Rheingold, 1993; Bruckman et al., 1994). When the internet commercialized in the 1990s, America OnLine offered perks to volunteer *community leaders* to manage its many chatrooms (Postigo, 2009). Today, volunteers continue to develop and carry out policies online, including Wikipedia (Forte, Larco, & Bruckman, 2009), Facebook (Facebook, n.d.), Twitter (Matias et al., 2015; R. S. Geiger, 2016) and Xbox (Good, 2013).

On reddit, volunteer moderators became the founders, facilitators, policymakers, and maintainers of communities with millions of subscribers, some of the largest conversations in the English language internet. Across the platform, reddit users came to expect that moderators would be impartial, unpaid, independent from the platform, and at least somewhat responsive to community demands (Matias, 2016a; Massanari, 2015). When the platform introduced features allowing participants to message moderators, designers deflected complaints about behavior away from the company and to these volunteers. As moderators tried to explain their decisions in response to increasing volumes of user complaint, many subreddits published statements with community policies.

In November 2012, the politics discussion group on reddit published a "frequently asked questions" (FAQ) document, which did not define harassment but noted that people who were being harassed could report it to the moderators (*reddit.com: help*, 2012). In mid-2013, moderators added the statement, "Please remember to observe proper reddiquette," near the bottom of their lengthy list of participation instructions. The idea of "rediquette" was promoted by the reddit company, who urged commenters to follow this "informal expression of the values of many redditors." The company encouraged people to "remember the human," "use proper grammar and spelling," and "don't be intentionally rude" (creesch, 2013).

By February 2014, the politics discussion community had reached 3 million subscribers. Moderators added an eleventh item to the sidebar, encouraging commenters to "please exercise civil discussion." They also replaced their FAQ with a "rules and regulations" document, which stated the community's first policies against hate speech and death threats (TheRedditPope, 2014). In the subsequent discussion, commenters expressed concerns about the risk of too many regulations. Several brought up civility as a category that could encom-

pass many issues and prevent the rules from growing longer.

What does it mean for a community to have policy? Researchers who study online community policymaking have tended to focus on Wikipedia, where policy documents play rhetorical or archival roles in ongoing deliberations over community governance (Butler, Joyce, & Pike, 2008). In a decentralized system like Wikipedia where any user can propose a new policy and any user can enforce it, policy documents are often attempts to convince others to do the work of governance in a certain way (Forte et al., 2009). From this perspective, the politics moderators on reddit became policymakers the day one of their moderators published a FAQ with participation guidelines. Yet governance on reddit is limited to groups of somewhat coordinated moderators, whose role is formally defined in the platform software. In r/politics, the community's first set of policy documents seem designed to disclaim responsibility rather than claim it, deferring even a definition of civility to the reddit company's broad message to "remember the human." Furthermore, formal rules rarely represent the beginning of a policy. As March argues in *The Dynamics of Rules*, rules look back as well as forward. They offer historical traces of the challenges faced by institutions and their responses to those challenges from the time before a rule was written (March, Schulz, & Zhou, 2000).

Scholars of policy evaluation tend to see policies as purposeful actions with intended outcomes, undertakings of consequence that can be discussed, decided, and evaluated (Heclo, 1972). Writing about content moderation, the legal scholar James Grimmelman imagines the components of a moderation policy as grammatical parts of a sentence. Rules about behavior are only the "subject" of a complete policy, which includes "verbs" of intervention and may also entail an "object" that the policymakers have in mind. In this view, even a notice encouraging commenters to "remember to observe proper reddiquette" is a policy. Moderators may have hoped that by including a civility reminder, commenters would behave more civilly, though the reminder was buried at the end of a longer list. Moderators may also have hoped that when people complained about removed comments, moderators could deflect those complaints by directing people to their public statements.

In August 2014, the politics discussion moderators made encouragements toward civility more prominent during "a bit of a makeover" (hoosakiwi, 2014). The encouragement to "be civil" was shifted from the least prominent to the most visible policy on the sidebar. The details were hidden, although curi-

ous users on desktop computers who chose to hover a mouse pointer over the guidelines and click see more detail (*Politics*, 2014; TheRedditPope, 2014). Moderators also modified the visual style of the comment form with a "CSS rule" that displayed a civility reminder to every person before they typed a new comment.

> Treat others with basic decency. No personal attacks, hate-speech, flaming, baiting, trolling, witch-hunting, or unsubstantiated accusations. Threats of violence will result in a ban

Policies like this civility warning include front-stage and backstage theories about the outcomes on behavior. By making the guidelines visible, moderators are working to influence social norms, commenters' perceptions about acceptable behavior in the community (Cialdini, Kallgren, & Reno, 1991). This front-stage message could influence the behavior of commenters whether or not moderators responded in any way to hate speech or consistently banned people for threats of violence. Any backstage actions by moderators to remove comments or ban users would involve backstage theories about the outcomes of those actions for the community and the people being sanctioned.

Most policies governing online behavior operate backstage and are never mentioned publicly. Instead they are evolving practices that moderators repeat, coordinate, and even encode in software. In interviews, moderators described varying levels of formality and coordination in their work. Few moderators go looking for rules to enforce. Instead moderators tend to participate frequently in their communities and intervene when they happen to see unacceptable behavior. Others monitor the "modqueue" of reports submitted by readers requesting that comments be removed. Moderators of large reddit communities often log into a community moderator chatroom while they work. When they are unsure about a decision, they ask other moderators who happen to be around. While many larger subreddits do consult communities, publish transparency reports, and hold moderator votes on major issues, the everyday work of defining acceptable behavior often occurs through this social feedback among moderators.

The most precise statements of a reddit community's policies are encoded in the software that controls a community's visual design and automated moderation systems. In 2014, the politics subreddit added style-sheet settings (CSS) to ensure that every commenter would see the civility reminder. Backstage, many

communities manage systematic decisions about which comments to remove using AutoModerator, a software agent empowered to make automated decisions based on lists of rules that can reach thousands of lines. The code behind these software systems is a kind of law, as Lessig puts it (Lessig, 2009), even if moderators must often tinker daily with those laws and carry out substantial work to reverse their own systems' governance mistakes.

Because policies in online communities are often defined and expressed in code, policy ideas are shaped, circulated, and negotiated by the stakeholders who define and adopt that code (Butler, Sproull, Kiesler, & Kraut, 2002). For example, before the politics subreddit could post civility reminders, the reddit company needed to implement software to allow communities to customize their appearance. Then, someone with software and design abilities needed to prototype the policy idea in a way that any community could adopt. Kelty describes this relationship between these stakeholders as an iterative, creative conversation among the "builders and imaginers" who craft platform and community software (Kelty, 2005). When communities do not possess the capacity to make these changes, they will sometimes advertise on moderation "job boards" and ask for technical help in policy discussion forums on the reddit site. When communities do find expert help, those experts also bring new policy ideas into communities along with technical support (Matias, 2016a).

Many community policy ideas cannot be implemented within the constraints of the a platform's software infrastructure. In 2014, one of the moderators of subreddit called "NotTheOnion" made a request for Automoderator civility reminders on the company's public forum for sharing "ideas for the admins." NotTheOnion is a subreddit for factual news that could be mistaken for satire. With 1.6 million subscribers at the time, moderators were struggling to manage conflict. Opposing political groups often disagreed over the amount of schadenfreude in the "politics and social or cultural issues" that appeared in the community, a moderator pointed out in an interview. At the time, this moderator was transitioning away from day-to-day moderation, was doing more to advise other communities on how to moderate, and was looking for work as a paid community manager for other platforms. When this NotTheOnion moderator noticed recurring patterns in "mob mentality" and "shit-flinging contests" across communities, they proposed that the company upgrade its software, allowing "mods to post a sticky comment to issue a reminder" at the top of a discussion. "This has been suggested forever" replied

another commenter in the discussion, which received no response from reddit employees. A search in the archive of ideas for the admins shows that moderators proposed the idea to the platform thirty-five times in the seven years before.

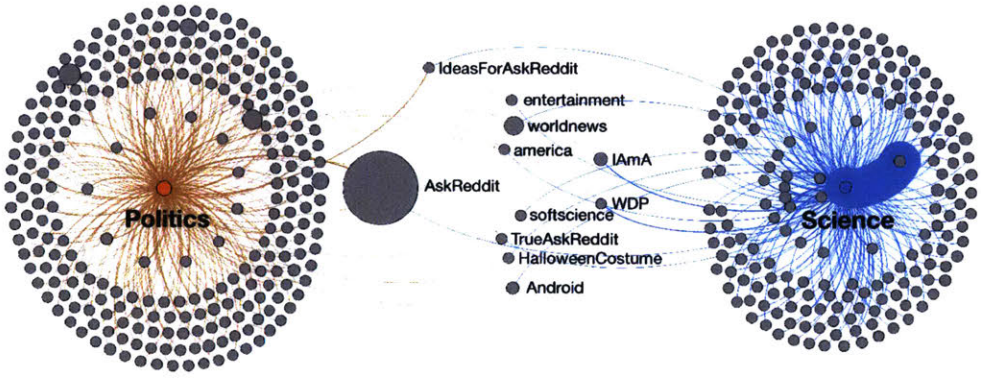


Figure 7-1: Moderators on reddit often take positions in more than one community, creating networks of shared governance that spread ideas about moderation. This graph shows a subset of the network of shared moderator positions in June 2015. Each node is a subreddit that shared at least one moderator with the politics or science subreddits (n=474). The area of a node is related to the number of comments a subreddit received that month, and edge thickness is determined by the number of shared moderators. In June 2015, the politics and science subreddits did not share any moderators, but some of their moderators likely interacted in other influential communities they moderated, like IamA and AskReddit. I created this chart with data from the population of moderator positions across 52,735 active subreddits that received at least one comment in June 2015. By the date r/science released experiment results in October 2016, the science and politics subreddits shared 3 common moderators. Shared moderation is only one network that predicts the co-occurrence of moderator actions. Other predictors include participation in moderation discussion groups and the relative insularity of commenters from the rest of the platform (Matias, 2016b).

By 2015, the relationship between community stakeholders and the reddit platform became more tense as the company struggled to manage groups that organized to mock and harass others. Moderators also felt pressure from systematic harassers and became less patient with the company's slow responses to their requests. On July 3, 2015, moderators of over 2,278 subreddits disabled their communities, "blacking out" in an attempt to force the company to respond to their expanded moderation software and closer coordination with the company. Within hours, the company agreed to meet moderator demands. In the months that followed, reddit hired new staff to implement platform features that expanded the policy capacity of community moderators. These sys-

tems improved moderator communication with their communities. The platform also provided the ability to post persistent "sticky comments" at the top of discussions. While the politics subreddit declined to join the blackout, their civility reminders relied on software that the platform created in response to other communities' advocacy (Matias, 2016b).

Figure 7-2: In April 2015, the reddit company released a mobile version that omitted all custom community design elements, including civility reminders. This change left communities searching for new ways to make their norms visible to commenters. The company officially launched its mobile web version in June 2016. This figure shows the desktop and mobile versions as they appeared in April 2017.

The moderator blackout in 2015 also illustrates the structure through which ideas about governance spread between communities on reddit. In research I previously published, I was able to show that the probability of a community to join the blackout could be predicted, in part, by factors in the structural relations of communities. Across the population of over fifty-two thousand active subreddits, I found that communities that share common moderators were likely to make similar decisions about participation in the blackout. Communities whose moderators also comment in discussions of moderation practices and platform-wide affairs were more likely to join the blackout. Furthermore, communities whose participants are more isolated from other communities on the platform were less likely to join the collective action (Matias, 2016b). These networks that enabled thousands of people on reddit to arrive at collective decisions about the strike against the company, may also spread policy information

(Figure 7-1).

When the reddit platform provided moderators with the ability to post persistent messages to the top of discussions, they were following their promises after the blackout to improve moderation tools. Company employees were also working to solve a problem that had been created by another change to the reddit system. In April 2015, the company began testing a mobile phone version of the site (Whitwam, 2015). The mobile phone interface disabled all community-specific design changes, effectively removing civility reminders displayed by the politics subreddit in the area where commenters type their message (Figure 7-2).

The politics subreddit did not adopt automated civility messages until ten months after the reddit platform made them available to communities in December 2015. Three factors converged to influence the community's decision: the U.S. presidential election, substantial use of the mobile site, and the adoption of civility messages by another community they admired. First, as the U.S. election became more prominent and rancorous, the demand increased for any intervention that could improve conversations and reduce moderator workloads. Second, the reddit company officially launched its mobile website in June 2016 (Amg137, 2016), preventing existing civility reminders from displaying for roughly half of unique commenters in some communities (Figure 7-2).[1] When a moderator noticed to the new sticky comment method being used in a smaller politics-related group, they proposed that the politics subreddit vote to do the same. With 11 of the group's moderators supporting, 4 opposing, and 3 abstaining, they approved the proposal and began to post automated civility reminders soon after.

While moderators of the politics subreddit did vote to adopt civility reminders during the 2016 presidential election, I have shown how the story of those reminders began much earlier. Across the ten year history of community governance on reddit, policy ideas develop and circulate through words and code alike. The path they travel is shaped by the design decisions of platform operators, the relationship networks of moderators, and negotiations between them. For a community to adopt a policy, the circumstances must present a demand for intervention and a community needs the the capacity to implement it. Evidence is only one of these many factors that inform an online community's

[1] private correspondence with reddit employee July 2016, March 2017

governance decisions.

Community Policy Evaluations

When community members and moderators of the politics subreddit referred to new evidence in their debates about civility reminders, they were having the kinds of discussions that originally motivated me to support community experiments. By supporting communities to conduct their own evaluations of community policies, I had hoped to see a growing network of online communities who shared and used the experimental knowledge that they create. For several months, I had worked with moderators of the science and world news subreddits to test the effects of similar notices. The politics subreddit offered early evidence that the results were spreading more widely.

Do messages stating expected norms of commenting behavior have any effect on people's actual commenting behavior? In February 2016 when moderators of the science subreddit suggested that we test the effects of automated rule postings, the feature had been available for two months and moderators were curious to find out. Over the next few months, I worked with moderators to develop a study design, developed software to carry out the field experiment. In late August and September, we conducted the experiment across 2218 discussions of scientific findings and live Q&A sessions with researchers. We found that posting the rules raised the chance that first-time comments would be permitted by moderators from 75.2% to 82.4%, a 7.3 percentage point increase on average in the science community (Matias, 2016c). On October 13, we held a community debriefing, reporting the results and fielding hundreds of questions from the community about the study and its implications (Matias, 2016d).

I supported a second community-led experiment from December 2016 through February 2017, after a science subreddit moderator introduced me to moderators from the world news discussion group. This community was struggling with community responses to sensationalized, misleading articles from unreliable news sources. Together, we tested the effect of encouraging fact-checking on the behavior of community commenters and the platform's algorithms (Matias, 2017a). We found that encouraging fact-checking did increase the rate at which commenters linked to further evidence in news discussions, and that by wording the encouragement to focus on human responses rather than algorithm outcomes, we could cause unreliable news to be demoted by the

reddit popularity algorithms. I shared the results in a community debriefing on Febuary 1, 2017, taking most of the day to respond to hundreds of community questions and ideas (Matias, 2017b).

My work to support communities policy experiments was inspired by the idea of an "experimenting society" proposed by policy evaluator Donald Campbell in 1971 (Campbell, 1998). In the essay, Campbell imagined networks of disputatious, local citizen experimenters who treat policy evaluation as another form of democratic participation. Just as we expect citizens to use their voices to influence social policies, Campbell imagined citizens using statistical analysis, experimentation, and replication in community decisions. The hallmark of this society, Campbell imagined, would be the ability of communities to reject evaluated policies, as well as debate policy ideas and conduct new experiments.

As online platforms become an intervention points for platform operators to govern a wide range of social problems, behavioral science research is shaping how people's daily lives are governed (S. Geiger, 2015), often without their consent (MacKinnon, 2012; Grimmelmann, 2015). When platforms delegate governance power to communities and their moderators, it becomes possible to create and evaluate policy with greater transparency, consent, and community participation. By supporting subreddit communities on reddit to conduct and discuss their own policy experiments, I hoped to take early steps toward an experimenting society, one where online communities create and share experimental evidence to support each other's self-governance.

In corporate data science, researchers tend to consider an experiment utilized if a decision to accept or reject an idea is based on that experiment (Regalado, 2014; White, 2012). In a democratic, experimenting society, Campbell imagined research as one part of a wider political process. He speculated that communities in an experimenting society might respond to new evaluations in a variety of ways:

- Experimental evidence would become part of community policy deliberation
- Citizens would question and re-analyze findings if supported through data literacy initiatives
- Communities would replicate each other's studies, cross-validating each other and determining the most appropriate and effective policies for

their context

Campbell's speech in 1971 was speculation, a "utopian" thought experiment he considered worth trying. In 2016, when I worked with the science subreddit to release our findings, we gained an opportunity to observe the diffusion of evidence from a community policy experiment across the population of communities on reddit. In the second part of this chapter, I share qualitative findings on the outcomes of adding experimental evidence to the words and code that shape policy on reddit.

How Experimental Evidence Informs Policymaking

Social researchers often hold mistaken beliefs that their work to evaluate a policy will lead the organization that commissioned their research to implement the idea, argued Weiss throughout a long scholarly career (Weiss, 1977). When Weiss first made this argument in the 1970s, a growing number of social scientists had been hired or commissioned by the U.S. government to evaluate social programs, and universities were launching policy schools to train new generations of policymakers and policy researchers (Heclo, 1972). Weiss observed that researchers expected a process where governments identify a problem, researchers generate knowledge to scope the problem, and further research would guide policymakers to choose among policy options (Figure 7-3) (Weiss, 1977).

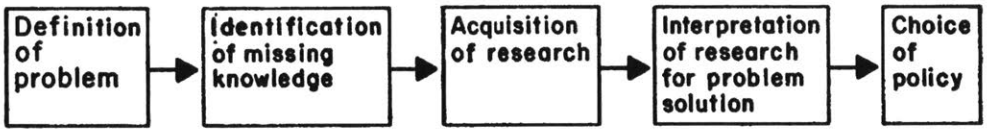


Figure 7-3: Many researchers mistakenly believe that is that research helps solve policy problems, according to Weiss (1977).

Research cannot help policymakers choose among well-defined options, Weiss argued, because policy decisions never occur in such a simple narrative. Organizations rarely meet formally to make large decisions at a specific moment in time. Instead, she argued, policies build up through ongoing patterns of practice that circumstances sometimes force to become formalized (Weiss, 1980). As Weiss continued to study the utilization of research over the next

thirty years, she came to argue that research influences policy through informal processes where research-informed ideas, inspiration, and ways of seeing are gradually appropriated by policymakers, who may not even be able to name their influences (Weiss, 1980). When research does reach policymakers, according to Weiss, it tends to serve four practical functions. Research often *legitimates* policymakers' pre-existing policy ideas. Other times, it *warns* policymakers about "conditions that are beyond the zone of acceptability." Findings sometimes provide *guidance* to policymakers, but that kind of influence can take decades. Finally, research provides *enlightenment* by crafting the "background of ideas, concepts, and information that increase their understanding of the policy terrain" (Weiss, 1995).

Within group decisions and democratic processes, research evidence is one of many resources that policy actors use to achieve goals and advance the interests they represent, according to Contandriopoulos and colleagues. Simplistic narratives about medical research often lead people into believing in direct, practical uses of research. In medicine, individual doctors are sometimes allowed to choose which treatments to employ. In similar communities with semi-independent practitioners, research evidence can influence practice quickly if many individuals adopt a well-packaged procedure. Unlike individual choices, group decisions require negotiation over more than evidence. Contandriopoulous's literature review found that the validity and strength of a study's causal results bear no relationship to its practical impact on policy set by groups (Contandriopoulos, Lemire, Denis, & Tremblay, 2010). Instead, the use of evidence in policy deliberation depends on the network structure of often-polarized political actors who seek to influence those deliberations (Contandriopoulos et al., 2010).

Researchers sometimes argue that participatory methods increase the chance that findings will lead directly to policy change. In principle, researchers can anticipate criticisms and deliberation by incorporating perspectives from multiple stakeholders in the design and interpretation of a study. Yet no single study can satisfy all interests equally (Cousins & Whitmore, 1998), a difficulty that increases with polarization in the structure of stakeholder relationships (Contandriopoulos et al., 2010).

Evidence from policy research often moves across networks rather than vertically between policymakers and the people they govern. In many cases, "policy entrepreneurs" advocate a governance idea across many communities,

hoping to develop a growing pool of evidence to influence even more widespread adoption (Mintrom, 1997). A substantial body of literature in political science works to predict the policy impacts of these advocates and the resources they marshal toward their goals, include research evidence (Contandriopoulos et al., 2010).

Because complex social and political factors influence the uses of evidence, policy evaluators have created institutions and infrastructures to generate and communicate research findings. Professional societies support evaluation in health, social policy, and international development, publishing meta-analyses of research findings on policy questions (Bero & Rennie, 1995; Chalmers, Hedges, & Cooper, 2002). In the UK, local What Works Centres consult for local government on social policy, creating a peer network for sharing evidence and replicating findings (Alexander & Letwin, 2013).

Any attempt on reddit to introduce evidence-based policy is likely to be shaped by similar forces to those observed in government policymaking. While the politics subreddit did hold a vote to adopt civility reminders in 2016, they also were formalizing a long-standing practice, something Weiss termed "decision accretion." Furthermore, the 2016 election put pressure on moderators to prioritize civility reminders. On reddit as in government, changes in the demand for a policy may determine when evidence is used. Furthermore, as I have shown, policies on reddit are advocated and spread through inter-community networks that sometimes collaborate and sometimes pressure the platform operators for change. When individual moderators support many communities, they sometimes work as policy entrepreneurs by sharing ideas between groups and advocating to the platform operators for new policy abilities.

Unlike other policy contexts, online platforms include their own built-in infrastructures for archiving and spreading ideas about governance. Throughout history, the content of policy discourse and the behavior of the embodied communities they govern have tended to remain categorically different, according to Kelty. Yet "recursive publics" like reddit "include not only the discourses of a public, but the ability to make, maintain, and manipulate the infrastructures of those discourses as well" (Kelty, 2005). Kelty imagines a conceptual spiral in which public discourse and policymaking influence each other: internet policies structure the trajectory of discourse, and that discourse structures the spread of policy in turn. Put another way, reddit communities face less friction than governments when sharing new evidence on the effects of policies,

Figure 7-4: **The Community Knowledge Spiral**: participatory processes that grow evidence-based policy in online communities.

since communities intrinsically include the means to find other policymakers, observe outcomes, and share evidence.

The Circulation and Uses of Evidence in Community Governance

When I developed field experiments with communities on reddit, I planned to explore the potential of a society where communities develop and share experimental evidence to govern the problems they face together. When planning these studies, I imagined a community knowledge spiral where each new study or replication prompts further discussion in a widening constellation of governance research (Figure 7-4). In many ways, this spiral resembles the myths of research impact that Weiss had thoroughly discredited. Yet tens of thousands of reddit communities already routinely circulate policy ideas and code through the infrastructures that host them. Like dust dropped into a river to observe its flow, I observed the uses of these first experiments to understand how communities might come to use their own and each other's research in an experimenting society.

To study how reddit communities make use of experimental knowledge, I interviewed a dozen moderators from communities that conducted experiments and others that discussed and used experiment findings. To find these communities, I recruited interview participants from moderator-focused discussion forums and through referrals from other moderators. I also analyzed

reddit-wide data on the use of automated civility reminders across over 180 communities, using methods of trace ethnography to guide further content analysis and interviews about moderation practices in those communities (R. S. Geiger & Ribes, 2011). To create this sample, I used a dataset of public reddit comments made in December 2016[2] to observe comments made by the AutoModerator, a common moderation system provided by the reddit platform that is capable of posting comments with rules to discussion threads. Since the AutoModerator makes public comments in the course of providing many other functions, I filtered the comments to only those that were direct replies to posts, as well as filtering out comments with words of direct address that are likely to be responses to individual users rather than general announcements. I then sent messages to those communities with requests for interviews. Participating moderators represented communities ranging from several thousand subscribers to over 15 million subscribers, on topics including news, science, politics, gaming, technology, gender, and image sharing.

Neither the science subreddit nor the world news subreddit made decisions about sticky comment policies after working for months to evaluate the effects of their interventions. In both cases, some moderators held informal discussions about the intervention and even discussed how they might adjust the policy. But by April 2017, neither community had reinstated governance practices that lapsed at the end of the experiment. As Weiss had found with government policymakers, neither community evaluation led to a decision of any kind among the alternatives that we had tested.

When I asked them why their communities might not have made the decision, moderators explained a mismatch between the low urgency for change and the substantial effort required to decide and implement the policies they had evaluated. A senior moderator in the science subreddit explained that implementing sticky comments would require new software development at a time when the community developers' limited resources were focused elsewhere. "It would require redoing a few of the bots," explained one of the moderators who maintains the community's software, and who remembered other moderators being "pleased and curious" about the results when we shared them with the community. In the science community, moderators also drew from their experience with statistical and experimental methods to interpret the re-

[2] data collected by Jason Baumgartner

sults. Pointing out the lower bound of the 95% confidence interval in the experiment results, one moderator argued in an interview that the intervention may not have much effect at all. If the confidence intervals had been tighter, this moderator argued, the study might have been more likely to prompt a group decision. In the world news community, moderators discussed variations on the policy they had evaluated. The moderators I interviewed expected that if the idea were proposed, they would likely support it unanimously. No proposal ever came to a decision, so no decision had been made.

While the results from community experiments did not lead communities to adopt the policies they tested, many moderators shared appreciation for the studies in public comments and private correspondence. Moderators called study results "insightful," "surprising," and "interesting." Several encouraged us to do more studies. As Weiss observed of late 20th century US policymakers, these moderators of participating communities seemed more interested in general enlightenment than in making a policy decision based on research (Weiss, 1979).

Beyond those that conducted the experiments, communities used research findings in decisions to adopt new policies and defend existing policies. When one comics discussion community began posting automated civility reminders in October, they linked directly to research findings in the message. After listing participation instructions, the civility reminder remarked that "/r/Science found that posting rules improves discussion quality." In the politics subreddit, as I have already described, commenters and moderators alike continued to link to the r/science study when responding to complaints about their civility reminders.

Some moderators in the science subreddit and elsewhere reported that experiment results convinced them to apply civility warnings selectively, as needed. When they anticipated conflict or saw a discussion with growing numbers of first-time commenters, they would post a reminder. "I personally brought it up several times [with other moderators] when we discussed new ways to improve the sub... nothing came to it," replied one moderator in an interview. Because the decision to post automated civility reminders required a high-effort voting process, this moderator and moderators of other groups decided to post reminders on a case-by-case basis, an action that only required approval in the moderator chatroom. In interviews, moderators sent me dozens of examples where they had taken individual initiative to encourage acceptable behavior.

Given that many policies in reddit communities build over time through the accretion of individual decisions, moderators were using the findings in the natural course of their governance work without group decisions.

Some moderators, working as policy entrepreneurs, used the experiment results to advocate for similar policies in other communities. In the world news community debriefing, a moderator from another subreddit asked, "How do you see this working [in my community] ?" Another wrote to me for more information on the studies, saying:

> I represent [two subreddits]. Both subs are looking for some automated reminders to help with rules, especially because of the recent election. I've been tasked with reaching out to the subs that currently use the reminders. [...] I know there's a few holdouts in the subs that doubt the effectiveness of a rule reminder and perhaps this would be a chance to convince them otherwise.

In discussions and interviews, other moderators were skeptical about differences between their communities and the ones that conducted experiments. "For a default subreddit or one the size of them like politics or news, absolutely," wrote one moderator. "In our community, absolutely not- our users would riot if they had to see that in every thread [laugh]." Advocates responded to these doubts by encouraging their peers to replicate findings with their own studies. If other moderators argued that the problems addressed by current evidence were unimportant to their community, these advocates encouraged their communities to find other policy ideas to evaluate. By April 2017, four new communities were planning community replications and three others were planning novel experiments.

Across communities, knowledge of experiment results was spread through moderator relationships and by platform algorithms. In October 2016 when the politics subreddit was debating civility reminders, a moderator of the science subreddit active in both communities was the first person to inform them of the results. The reddit platform's popularity algorithms also spread the results widely. In October, the science subreddit's community debriefing received enough activity that the platform's algorithms promoted the discussion to the front page of reddit for 30 hours, and the results were viewed over 240,000 readers. In 2017 when I reported results to the world news community, some-

one else linked to the conversation from a popular group for sharing the "best of reddit," a link that remained on the front page for six hours.[3]

In February 2017, the reddit platform announced that they planned to study the effects of showing community rules at the top of discussions (powerlanguage, 2017). When employees asked for volunteers, over a hundred communities joined, including many of the largest communities on the platform. When I asked employees if our community experiments had influenced their work, the lead designer responded, "I was definitely aware of your study." The designer described details of the science subreddit's experiment and outlined further questions that the company hoped to answer with this new research. Instead of showing civility reminders to all readers, which led frequent commenters to complain, the designers planned to test the effects of showing civility reminders only to new or infrequent readers. Evidence from community-led experiments had reached the platform designers, shaping policy research in over a hundred other communities and potentially leading to fundamental changes in the reddit platform.

While moderators' uses and non-uses of policy evidence resemble similar patterns among government policymakers, the characteristics of the reddit platform enabled community research to influence over a hundred other communities in just four months. Even if communities that commissioned the studies didn't make decisions on the findings, other communities who felt the need more urgently used the results to defend existing policies and adopt new ones. Some moderators adopted the policies in their personal moderation work, without group decisions. Other moderators acted as policy entrepreneurs, advocating for the use of policy evidence across communities. When communities expressed uncertainty, these advocates worked to convince them to conduct replications or new studies. Many of these exchanges were mediated by the software design of the reddit platform, where public conversation combined with popularity algorithms to spread research findings to hundreds of thousands of people. Research also influenced policy work in over a hundred communities after platform designers responded with new features and study replications.

[3] I estimated the time on the front page by querying the front page of reddit for the top 100 items every 4 minutes during the community debriefings.

Lessons for An Experimenting Society

In 2016 when moderators of the politics community on reddit intervened with civility reminders during the height of conflict over the U.S. election, they were adopting policies that shaped the political discourse of millions of people. In the kind of experimenting society imagined by Donald Campbell, this would be common. He argued that communities could routinely conduct new policy evaluations and learn from each other in a disputatious network of citizen policy evaluators (Campbell, 1998). For these policy experiments to become a routine civic action, communities need to circulate, debate, use, and reject research knowledge as well as create it.

By supporting communities on reddit to conduct their own policy experiments, I was able to release new evidence into the currents of community policymaking and observe how that evidence was used. I end this chapter with lessons for evidence-based community governance based on those observations.

To flourish, an experimenting society needs to overcome the problem that communities who most need evidence are often the least likely to create that evidence. Communities often introduce new policies in response to the demand from growing problems. Those communities may be less likely to conduct new research than more stable communities, especially if a study will take months to complete. Several strategies might serve communities more effectively. First, methodologists can make technical advancements in policy evaluation methods by reducing the time to conduct a study. Second, researchers could focus on conducting experiments with communities that are in a position to test ideas, producing evaluations that could guide more beseiged moderators. Third, researchers could expand the available evidence by developing systems to generate quasi-experimental findings for communities about historical policy decisions. Finally, communities could adopt contextual bandit systems that adaptively-select the optimal intervention over time, and perhaps extract generalizable knowledge from the data collected by those bandits (White, 2012).

Researchers who wish to support policy networks to learn together through experiments can prioritize projects based on the shape of those networks. Researchers could prioritize policy questions of interest to many communities and begin with well-known communities that have strong ties elsewhere. Moderators, participants, and observers might carry what they learn across other

communities, who might adopt those policies or develop new studies.

Public community debriefings spread knowledge about research findings while also performing an important role in community consent. On reddit, debriefings can become impromptu spaces for other communities to discuss policies and imagine new studies. When popularity algorithms spread a debriefing conversation beyond the community that hosted a study, research findings can become widely known.

Because platform policies and experiments involve software as well as human activity, code is one of the greatest barriers to policy adoption and one of the greatest means to spread an idea. The science subreddit delayed adopting the policy they tested due to the effort required to implement it. Yet when reddit designers implemented similar software and invited communities to test it, over a hundred offered to participate in the company experiment. If researchers package new studies as easy-to-deploy policy intervention software, they could expand the number of evaluations and increase the use of research by communities. Finally, the widest influence from community-led experiments may occur when communities answer questions that platform operators are also asking, since changes by employees shape the policy capacities of every community across a platform.

Across the internet, volunteer moderators and bystanders carry out a substantial role to govern the social interactions of hundreds of millions of people. By supporting these policymakers to evaluate the outcomes of their work, we add evidence to moderators' policy discussions with each other and the communities they serve. As communities debate their values, experimental knowledge can help them achieve the values they agree on together. Yet evidence is only part of these deliberative activities. Endeavors toward an experimenting society must extend beyond research methods. We should apply equally-creative effort into the diffusion and uses of policy evidence by online communities.

References

Alexander, D., & Letwin, O. (2013). What Works: evidence centres for social policy. Retrieved 2017-03-26, from `http://dera.ioe.ac.uk/17396/1/What_Works_publication.pdf`

Amg137. (2016, June). *New look on Reddit mobile web: compact view -*

r/announcements. Retrieved 2017-04-08, from https://www.reddit.com/r/announcements/comments/4nc81l/new_look_on_reddit_mobile_web_compact_view/

Barthel, M., Stocking, G., Holcomb, J., & Mitchell, A. (2016, February). *Seven-in-Ten Reddit Users Get News on the Site* (Tech. Rep.). Pew Research Center for Journalism & Media. Retrieved 2017-04-02, from http://www.journalism.org/2016/02/25/seven-in-ten-reddit-users-get-news-on-the-site/

Bero, L., & Rennie, D. (1995). The Cochrane Collaboration: preparing, maintaining, and disseminating systematic reviews of the effects of health care. *Jama, 274*(24), 1935–1938. Retrieved 2017-03-26, from http://jama.jamanetwork.com/article.aspx?articleid=393319

Bruckman, A., Curtis, P., Figallo, C., & Laurel, B. (1994). Approaches to managing deviant behavior in virtual communities. In *CHI Conference Companion* (pp. 183–184).

Butler, B., Joyce, E., & Pike, J. (2008). Don't look now, but we've created a bureaucracy: the nature and roles of policies and rules in wikipedia. In *Proceedings of the SIGCHI conference on human factors in computing systems* (pp. 1101–1110). ACM. Retrieved 2017-04-06, from http://dl.acm.org/citation.cfm?id=1357227

Butler, B., Sproull, L., Kiesler, S., & Kraut, R. (2002). Community effort in online groups: Who does the work and why. *Leadership at a distance: Research in technologically supported work*, 171–194.

Campbell, D. T. (1998). The experimenting society. In *The experimenting society: Essays in honor of Donald T. Campbell* (p. 35). New Brunswick: Transaction Publishers.

Chalmers, I., Hedges, L. V., & Cooper, H. (2002, March). A Brief History of Research Synthesis. *Evaluation & the Health Professions, 25*(1), 12–37. Retrieved 2017-03-26, from http://dx.doi.org/10.1177/0163278702025001003 doi: 10.1177/0163278702025001003

Cialdini, R. B., Kallgren, C. A., & Reno, R. R. (1991). 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.

Contandriopoulos, D., Lemire, M., Denis, J.-L., & Tremblay, . (2010, December). Knowledge Exchange Processes in Organizations and Policy Arenas: A Narrative Systematic Review of the Literature. *Milbank Quarterly, 88*(4), 444–483. Retrieved 2017-03-17, from http://onlinelibrary.wiley.com.libproxy.mit.edu/doi/10.1111/j.1468-0009.2010.00608.x/abstract doi: 10.1111/j.1468-0009.2010.00608.x

Cousins, J. B., & Whitmore, E. (1998). Framing participatory evaluation. *New directions for evaluation, 1998*(80), 5–23. Retrieved 2017-03-17, from http://onlinelibrary.wiley.com/doi/10.1002/ev.1114/full

creesch. (2013, January). *reddit: the front page of the internet.* Retrieved 2017-04-08, from http://web.archive.org/web/20130126111425/http://www.reddit.com/wiki/reddiquette

Facebook. (n.d.). *Group Admin Basics: What is a Group Admin?* Retrieved 2017-03-29, from https://www.facebook.com/help/418065968237061/

Forte, A., Larco, V., & Bruckman, A. (2009). Decentralization in Wikipedia governance. *Journal of Management Information Systems, 26*(1), 49–72. Retrieved 2017-04-06, from http://www.tandfonline.com/doi/abs/10.2753/MIS0742-1222260103

Geiger, R. S. (2016, June). Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society, 19*(6), 787–803. Retrieved 2016-08-29, from http://dx.doi.org/10.1080/1369118X.2016.1153700 doi: 10.1080/1369118X.2016.1153700

Geiger, R. S., & Ribes, D. (2011). Trace ethnography: Following coordination through documentary practices. In *System Sciences (HICSS), 2011 44th Hawaii International Conference on* (pp. 1–10). IEEE. Retrieved 2017-04-03, from http://ieeexplore.ieee.org/abstract/document/5718606/

Geiger, S. (2015). Does facebook have civil servants? On governmentality and computational social science. In *Workshop on Ethics for Studying Sociotech-*

nical Systems in a Big Data World. Vancouver, British Columbia, Canada. Retrieved from https://cscwethics2015.files.wordpress.com/2015/02/geiger.pdf

Good, O. (2013, August). Does Your Gamertag Have Herpes? Beware Xbox Live Enforcement United. *Kotaku.* Retrieved 2015-09-23, from http://kotaku.com/does-your-gamertag-have-herpes-beware-xbox-live-enfor-1019141385

Gottfried, J., Barthel, M., & Mitchell, A. (2017, January). *Trump, Clinton Voters Divided in Their Main Source for Election News.* Retrieved 2017-04-02, from http://www.journalism.org/2017/01/18/trump-clinton-voters-divided-in-their-main-source-for-election-news/

Grimmelmann, J. (2015). The law and ethics of experiments on social media users. Retrieved 2017-03-28, from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2604168

Heclo, H. H. (1972). Review article: Policy analysis. *British journal of political science, 2*(01), 83–108. Retrieved 2017-03-31, from http://journals.cambridge.org/article_S0007123400008449

hoosakiwi. (2014, July). *A new sidebar and a wiki update to match. r/politics.* Retrieved 2017-04-08, from https://www.reddit.com/r/politics/comments/2bm486/a_new_sidebar_and_a_wiki_update_to_match/

Kelty, C. (2005). Geeks, social imaginaries, and recursive publics. *Cultural Anthropology, 20*(2), 185–214. Retrieved 2017-04-06, from http://onlinelibrary.wiley.com/doi/10.1525/can.2005.20.2.185/full

Lessig, L. (2009). *Code: And other laws of cyberspace.* Basic Books.

MacKinnon, R. (2012). *Consent of the Networked: The Worldwide Struggle for Internet Freedom.* Basic Books.

March, J. G., Schulz, M., & Zhou, X. (2000). *The dynamics of rules: Change in written organizational codes.* Stanford University Press.

Margetts, H., John, P., Hale, S., & Yasseri, T. (2015). *Political turbulence: how social media shape collective action.* Princeton University Press.

Massanari, A. L. (2015). *Participatory Culture, Community, and Play: Learning from Reddit* (1st New edition edition ed.). New York: Peter Lang Inc., International Academic Publishers.

Matias, J. N. (2016a, September). The Civic Labor of Online Moderators. Oxford, UK. Retrieved 2017-03-29, from http://ipp.oii.ox.ac.uk/sites/ipp/files/documents/JNM-The_Civic_Labor_of_Online_Moderators__Internet_Politics_Policy_.pdf

Matias, J. N. (2016b). Going Dark: Social Factors in Collective Action Against Platform Operators in the Reddit Blackout. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (pp. 1138–1151). ACM. Retrieved 2017-03-29, from http://dl.acm.org/citation.cfm?id=2858391

Matias, J. N. (2016c, October). *Posting Rules in Online Discussions Prevents Problems & Increases Participation.* Retrieved 2017-03-27, from http://civilservant.io/r_science_sticky_coments_1.html

Matias, J. N. (2016d, October). *Posting Rules in Online Discussions Prevents Problems & Increases Participation, in a Field Experiment of 2,214 Discussions On r/science - r/science.* Retrieved 2017-04-08, from https://www.reddit.com/r/science/comments/56h704/posting_rules_in_online_discussions_prevents/

Matias, J. N. (2017a, February). *Persuading Algorithms With an AI Nudge.* Retrieved 2017-03-27, from https://civilservant.io/persuading_ais_preserving_liberties_r_worldnews.html

Matias, J. N. (2017b, February). *Sticky Comments Increase Fact-Checking and Cause Tabloid News To Be Featured Less Prominently on reddit : worldnews.* Retrieved 2017-04-10, from https://www.reddit.com/r/worldnews/comments/5rg40h/sticky_comments_increase_factchecking_and_cause/

Matias, J. N., Johnson, A., Boesel, W. E., Keegan, B., Friedman, J., & DeTar, C. (2015). Reporting, Reviewing, and Responding to Harassment on Twitter. *arXiv preprint arXiv:1505.03359.* Retrieved 2015-11-10, from http://arxiv.org/abs/1505.03359

Menegus, B. (2016, January). Reddit Is Tearing Itself Apart. *Gizmodo*. Retrieved 2017-04-02, from http://gizmodo.com/reddit-is-tearing-itself-apart-1789406294

Mintrom, M. (1997). Policy entrepreneurs and the diffusion of innovation. *American journal of political science*, 738–770. Retrieved 2017-03-31, from http://www.jstor.org/stable/2111674

Phillips, W., Beyer, J., & Coleman, G. (2017, March). Trolling Scholars Debunk the Idea That the Alt-Right's Shitposters Have Magic Powers. *Vice*. Retrieved 2017-04-02, from https://motherboard.vice.com/en_us/article/trolling-scholars-debunk-the-idea-that-the-alt-rights-trolls-have-magic-powers

Pictures and Images. (2009, September). Retrieved 2017-04-08, from http://web.archive.org/web/20090909235635/http://www.reddit.com/r/pics?

Politics. (2014, February). Retrieved 2017-04-08, from http://web.archive.org/web/20140202125743/http://www.reddit.com/r/politics/

Postigo, H. (2009, September). America Online volunteers. *International Journal of Cultural Studies*, *12*(5), 451–469. Retrieved 2015-08-19, from http://ics.sagepub.com.libproxy.mit.edu/content/12/5/451

powerlanguage. (2017, February). *Improvements to subreddit rules - r/modnews.* Retrieved 2017-04-10, from https://www.reddit.com/r/modnews/comments/5u9yh8/improvements_to_subreddit_rules/

reddit.com: help. (2012, November). Retrieved 2017-04-08, from http://web.archive.org/web/20121128142404/http://www.reddit.com/help/faqs/politics

Regalado, A. (2014, January). Websites Turn to Experiments. *MIT Technology Review*. Retrieved 2017-04-01, from https://www.technologyreview.com/s/523671/seeking-edge-websites-turn-to-experiments/

Rheingold, H. (1993). *The virtual community: Homesteading on the electronic frontier.* MIT press.

Singal, J. (2016, September). How Internet Trolls Won the 2016 Presidential Election. *New York Magazine.* Retrieved 2017-04-02, from http://nymag.com/selectall/2016/09/how-internet-trolls-won-the-2016-presidential-election.html

TheRedditPope. (2014, February). *rulesandregs - politics.* Retrieved 2017-04-08, from https://www.reddit.com/r/politics/wiki/rulesandregs?v=2c513958-0c69-11e3-b89d-12313d1940ac

Weiss, C. H. (1977). Research for policy's sake: The enlightenment function of social research. *Policy analysis,* 531–545. Retrieved 2017-03-17, from http://www.jstor.org/stable/42783234

Weiss, C. H. (1979). The many meanings of research utilization. *Public administration review, 39*(5), 426–431. Retrieved 2017-03-17, from http://www.jstor.org/stable/3109916

Weiss, C. H. (1980). Knowledge creep and decision accretion. *Knowledge, 1*(3), 381–404. Retrieved 2017-03-17, from http://journals.sagepub.com/doi/pdf/10.1177/107554708000100303

Weiss, C. H. (1995). The haphazard connection: social science and public policy. *International Journal of Educational Research, 23*(2), 137–150. Retrieved 2017-03-31, from http://www.sciencedirect.com/science/article/pii/0883035595914986

White, J. (2012). *Bandit algorithms for website optimization.* O'Reilly Media, Inc.

Whitwam, R. (2015, April). *Reddit Has A New Beta Mobile Site, And It Doesn't Suck.* Retrieved 2017-04-08, from http://www.androidpolice.com/2015/04/21/reddit-has-a-new-beta-mobile-site-and-it-doesnt-suck/

Williams, A. (2016, October). How Pepe the Frog and Nasty Woman Are Shaping the Election. *The New York Times.* Retrieved 2017-04-02, from https://www.nytimes.com/2016/10/30/style/know-your-meme-pepe-the-frog-nasty-woman-presidential-election.html

Chapter8

Epilogue

Designing the Experimenting Society

Across this dissertation, I have searched for ways to reconcile the tremendous power of platform governance with the need to use that power wisely through democratic means.

In the opening paper, I drew inspiration from the history of struggles to redesign social experiments as civic participation in an open, experimenting society. In my fieldwork with community moderators on reddit, I reported the kinds of accountability required of their civic labor. Choosing to continue working with reddit communities, I designed CivilServant to help them evaluate their governance work with community-led policy experiments. In the findings I have reported, communities demonstrated tremendous creativity as we tested the effects of their policies on human and machine behavior. Finally, since evidence cannot serve communities without discussion and action. I have followed the trail of community evidence we created together.I found that because platforms are designed to spread discourse and code, evidence from experiments can circulate rapidly to influence policy, platform design, and further community research.

As I write this, seven communities on reddit are considering new policy experiments. Moderators on other large platforms have asked for the ability to do the same. The CivilServant project is launching shortly as a nonprofit and looks likely to increase the rate of community policy experiments. These developments leave me thinking about the future of platform governance and the role that communities might play in that future.

As a designer, I am profoundly aware of my limitations to create or even define the conditions of a just, experimenting society. Each community evaluation could legitimately be criticized over the relative power of stakeholders and

the the degree to which the subsequent discussion resembled what a given critic considers to be legitimate democracy. Previous attempts at collective knowledge online have been plagued by inequality and other problems that spring from the cultures where they began. Community-led policy evaluation could follow the same path. In other words, the experimenting society faces all of the problems of democracy, where structures of oppression or differences in material conditions, culture, and literacies influence the contours of power and justice.

When we choose an open society, we hope that the outcome of collective arguments and compromises will benefit society, or at least as Popper argued, allow society to reject the worst abuses of power. At a time when platform governance is rarely guided by research, mostly conducted in secret, and infrequently open to public deliberation, community-led experiments create new opportunities for those public arguments and compromises.

In April 2017, a New York Times Magazine article called online harassment a pervasive problem which might be "a lost cause" (Wortham, 2017). Platforms offer a powerful point of intervention on social problems, but social change moves slowly. In the U.S. today, 47% of internet users have experienced some kind of harassment, with 27% of Americans self-censoring their online behavior out of fear of harassment (Lenhart, Ybarra, Zickuhr, & Price-Feeney, 2016). Even if a one-percentage-point change could improve the lives of millions of people, these problems would still remain. Yet we have good evidence for hope with online harassment, the area that first motivated me to study how platforms can govern behavior. Of the 72% of U.S. internet users who have witnessed harassment, 65% reported taking some action to intervene or support the harassment receiver (Lenhart et al., 2016). With each new study we do together, we empower each other to make more informed choices about how address enduring problems like harassment.

Across my work to develop this dissertation, I have learned need for enduring hope and action on problems that yield to our collective work in the long term. These early results make me hopeful for the future of governing human and machine behavior in experimenting societies.

References

Lenhart, A., Ybarra, M., Zickuhr, K., & Price-Feeney, M. (2016, November).

Online Harassment, Digital Abuse, and Cyberstalking in America (Tech. Rep.). Data & Society Research Institute.

Wortham, J. (2017, April). Why Can't Silicon Valley Fix Online Harassment? *The New York Times Magazine.* Retrieved 2017-04-16, from https://www.nytimes.com/2017/04/04/magazine/why-cant-silicon-valley-fix-online-harassment.html

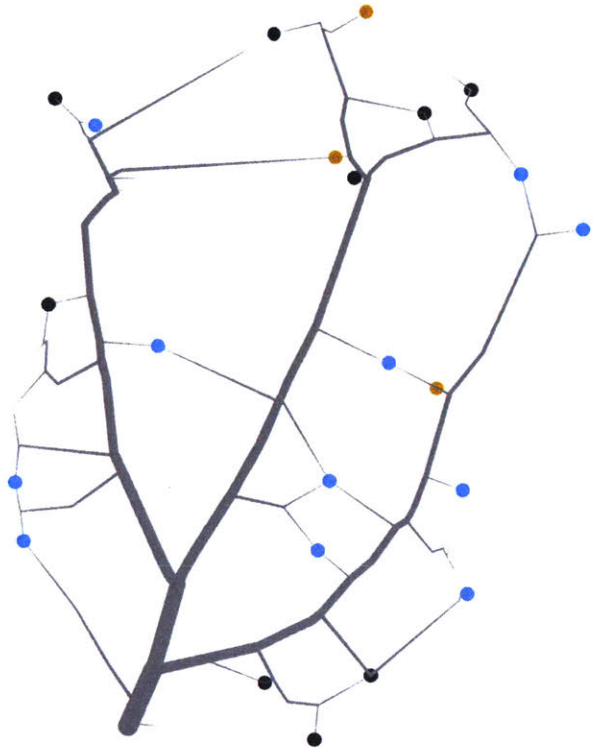
AppendixA

Illustrating Average Treatment Effects in Community-Led Experiments

Throughout this dissertation, I found myself frustrated that community participants were supporting a remarkable collaborative endeavor by joining experiments and would receive a single bar chart for their efforts. While the art of information visualization has produced beautiful, inspiring forms of correlational data, the same has not been true for causal inference.

To thank participants and express the values of their contribution beyond the numbers, I drew inspiration from procedural, generative algorithms for producing organic forms. Based on leaf venation algorithms (Runions et al., 2005) and open source software by Anders Hoff (Hoff, n.d.), I created thank-you cards that simulated the output of the r/science experiment's statistical mode for specific conversations. On one side, these cards show details of a control-group conversation as it occurred. On the other side, I have simulated four possible conversations that might have happened if the anticipated average treatment effect were to apply. While the actual structure of comments and replies is simulated, each dot represents a comment that might have been made, and whether it might have been removed by moderators in the r/science experiment.

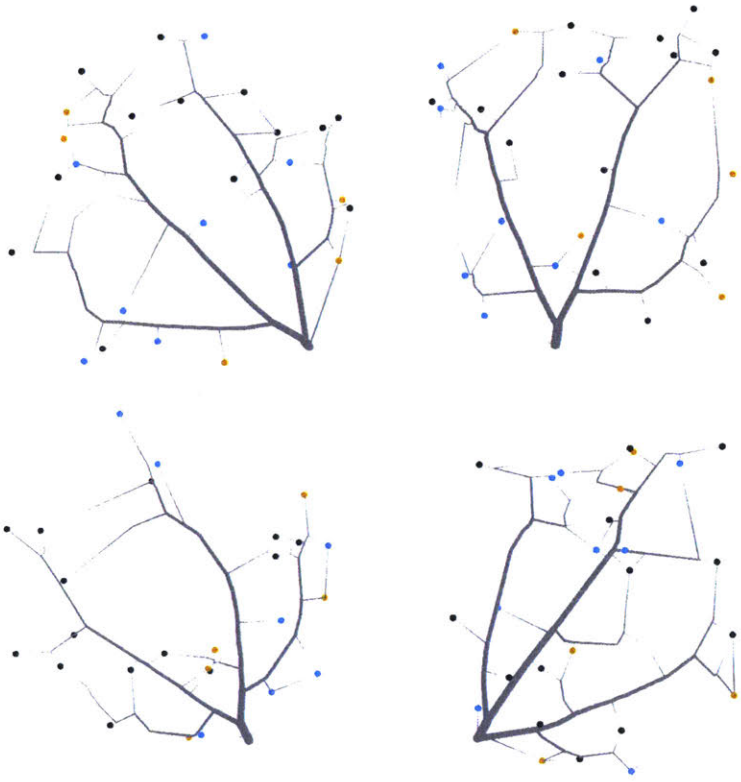
These early prototypes represents a first effort toward re-imagining how we illustrate community-led policy experiments to embody the wider values of community-building these experiments support.



Autism discussion in r/science on reddit, September 2016
Simulated leaf patterns are based on comments from this conversation, which did not display community rules

• 24 comments • 8 newcomers • 5 first comments removed

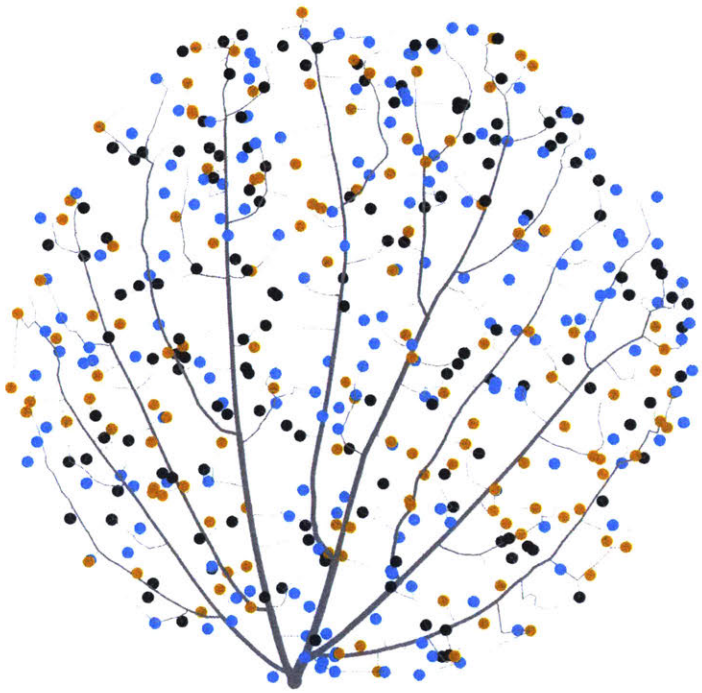
Experiment details are at **bit.ly/cs-science-2016**
(discussion structure is simulated to protect participant privacy)



Four possible autism discussions if rules had been displayed

- 38.1% increase in first-time comments by newcomers
- 56% newcomer comments removed rather than 63%
- 27 comments • 11 newcomers • 6 first comments removed

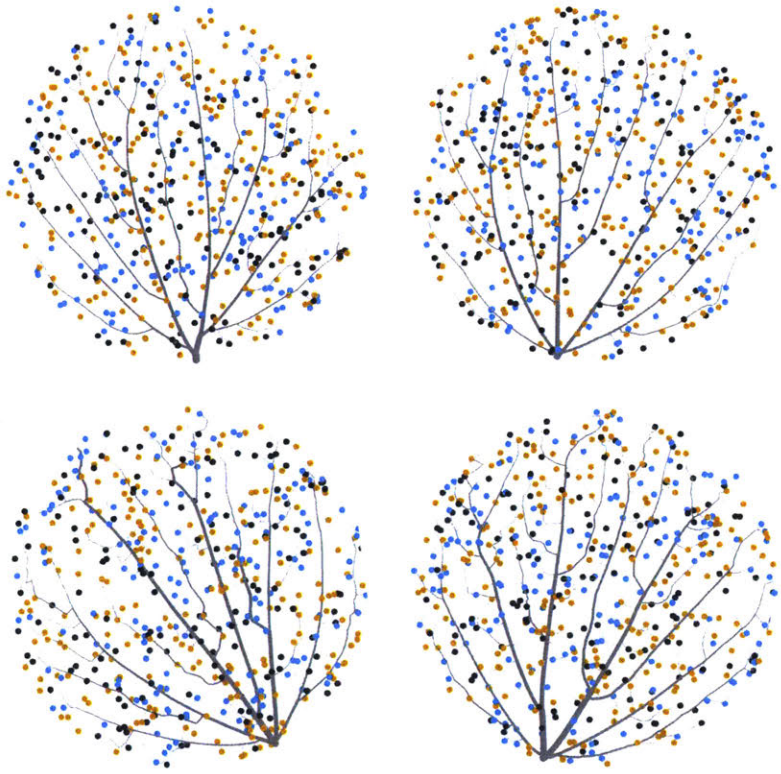
J. Nathan Matias (@natematias) **civilservant.io**



Zika & pregnancy discussion in r/science on reddit, September 2016
Simulated leaf patterns are based on comments from this conversation, which did not display community rules

• 455 comments • 179 newcomers • 51 first comments removed

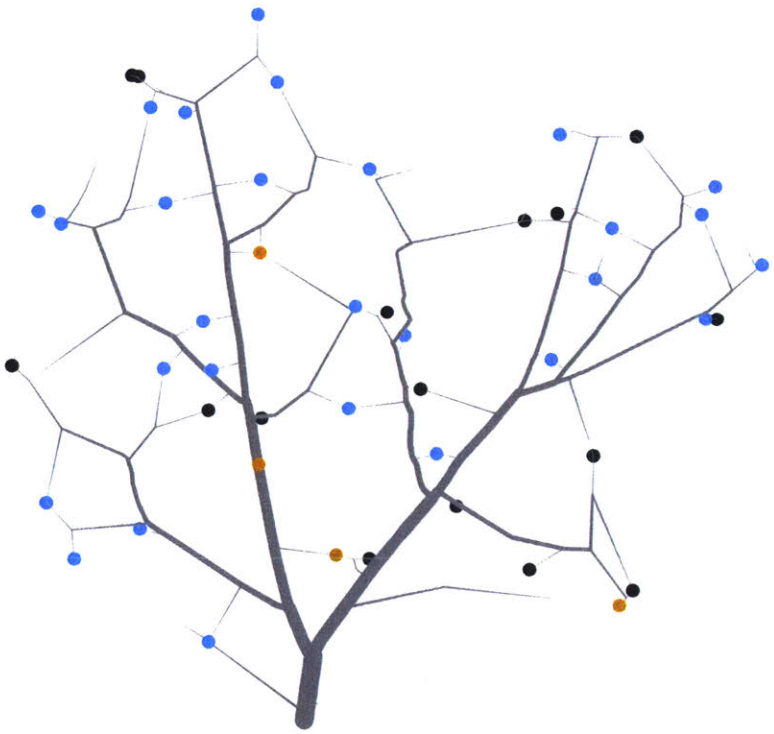
Experiment details are at **bit.ly/cs-science-2016**
(discussion structure is simulated to protect participant privacy)



Four possible Zika & pregnancy discussions if rules had been displayed

- 38.1% increase in first-time comments by newcomers
- 21% newcomer comments removed rather than 28%
- 523 comments • 247 newcomers • 38 first comments removed

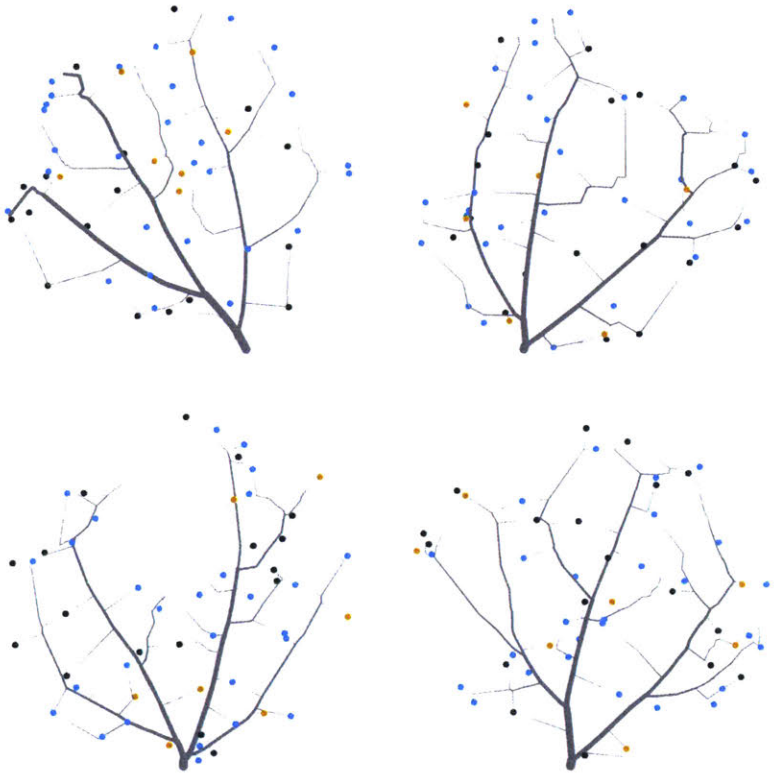
J. Nathan Matias (@natematias) **civilservant.io**



Insect discovery discussion in r/science on reddit, September 2016
Simulated leaf patterns are based on comments from this conversation, which did not display community rules

• 49 comments • 10 newcomers • 6 first comments removed

Experiment details are at **bit.ly/cs-science-2016**
(discussion structure is simulated to protect participant privacy)



Four possible insect discovery discussions if rules had been displayed

- 38.1% increase in first-time comments by newcomers
- 53% newcomer comments removed rather than 60%
- 54 comments • 14 newcomers • 7 first comments removed

J. Nathan Matias (@natematias) **civilservant.io**

References

Hoff, I., Anders. (n.d.). *On Generative Algorithms.* Retrieved 2017-05-14, from `http://inconvergent.net/generative/`

Runions, A., Fuhrer, M., Lane, B., Federl, P., Rolland-Lagan, A.-G., & Prusinkiewicz, P. (2005). Modeling and visualization of leaf venation patterns. *ACM Transactions on Graphics (TOG), 24*(3), 702–711. Retrieved 2017-05-14, from `http://dl.acm.org/citation.cfm?id=1073251`