CivilServant: Community-Led Experiments in Platform Governance

J. Nathan Matias
Massachusetts Institute of Technology

Merry Mou
Massachusetts Institute of Technology

ABSTRACT

As online platforms monitor and intervene in the daily lives of billions of people, platforms are being used to govern enduring social problems. Field experiments could inform wise uses of this power if tensions between democratic values and experimentation could be resolved. In this paper, we introduce CivilServant, a novel experimentation infrastructure that online communities and their moderators use to evaluate policies and replicate each others' findings. We situate CivilServant in the political history of policy experiments and present design considerations for community participation, ethics, and replication. Based on two case studies of community-led experiments and public debriefings on the reddit platform, we share findings on community deliberation about experiment results. We also report on uses of evidence, finding that experiments informed moderator practices, community policies, and replications by communities and platforms. We discuss the implications of these findings for evaluating platform governance in an open, democratic, experimenting society.

Author Keywords

governance, moderation, field experiments, randomized trials, action research, ethics, platforms, policy evaluation

ACM Classification Keywords

K.4.1 Computers and Society: Public Policy Issues: Use and Abuse of Power

INTRODUCTION

As social platforms and intelligent agents become routine in the daily life of billions of people, the public has come to expect these systems to address deep-seated social ills. Platforms are currently expected to manage social problems including terrorism [49, 75], discrimination [59, 23, 24], suicide [55], self-harm [14], eating disorders [14], hate speech [17], child pornography [73], misogyny [4], copyright violation [70], and political polarization [71], to name a few. In recent years, mainstream advocacy organizations have established branches in San Francisco, hoping to influence U.S. platforms to adopt policies on issues usually addressed by legislation [13]. Even as platforms attempt to retain a non-intervention stance [33], their designers and researchers have arguably become powerful civil servants who govern human affairs [31].

This pressure on platforms relies on an assumption that platforms possess effective means to govern society. Yet well-intentioned policies and designs have sometimes increased a problem or caused disastrous side-effects for years before the effects were known [15, 36]. Just as evaluations have not matched the rate of new interventions, the diversity of evaluations has not scaled to match the range of human culture that platforms govern [38]. Furthermore, many findings remain secret within companies that are incentivized to protect their reputations and their intellectual property [53]. Consequently, public assumptions about the benefits of platform interventions remain unproven while failures go unnoticed.

Endeavors to govern social behavior through platforms take one of two strategies. The most common strategy requires platform operators to take governance power, build policy teams, and hopefully develop the research capacities to evaluate those policies [17, 33, 31, 8]. Advocates and governments then attempt to hold platforms accountable for their use of governance power [50]. A second strategy delegates power from platforms to civil society organizations and volunteer moderators, who then create and enact their own local policies [64, 35, 16, 66]. While the civic labor of this delegated governance can be difficult to sustain [51], delegation can scale governance work, adapt to cultural differences, and make public accountability a civic process rather than a commercial process [30]. People who hold delegated power can sometimes be more accountable to the people they govern, yet they almost never have the capacity to evaluate the outcomes of their work to regulate hate speech, manage conflict, enforce copyright laws, or govern public discourse.

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

CivilServant participates in a history of debates on the role of social experiments in democratic societies. We situate CivilServant within that history, offer design considerations for community-led experiment infrastructures, describe the system, and summarize the research process. We illustrate the system's uses with two case studies, policy evaluations by

communities with over 12 million participants each. For each case, we report community deliberation and uses of research findings. We conclude with challenges 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" [63, 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" [63, 107]. For Popper, 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 experiments were more common, writes Popper, "politicians [might] begin to look out for their own mistakes instead of trying to explain them away and to prove that they have always been right." Yet closed society policymakers, who Popper calls "utopian engineers," evaluate policies and shape society without regard for citizen views [63, 143].

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" [42, 58]. As the U.S. government adopted research methods from Campbell's textbooks [11], 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 [10]. 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 [11, 49]. Campbell challenged methodologists to redesign their methods to include "individual participation and consent at all decision levels possible" [10, 42]. At a time when field experiments were rare, he imagined a society where local communities conducted plentiful policy studies across a network 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" [10].

By advocating for democratic networks of replication and cross-validation, Campbell anticipated later developments in feminist standpoint theory that grounded empirical research in the position and perspectives of communities, according to the feminist sociologist Anne Oakley [58]. In Campbell's original speech, he hoped that faculty at small regional U.S. colleges would be funded to conduct community experiments and replications with local governments [9]. Instead, Campbell's proposal remained a thought experiment distributed and debated by practitioners in the policy evaluation field [58].

Many HCI researchers use methods that collaborate with community partners, in an action research tradition [37]. Campbell's experimenting society differs from this tradition in two ways: its use of randomized trials and its statistical vision for the circulation of community-generated knowledge.

Despite the origin of action research in factory experiments [1], some action researchers in HCI see randomized trials as fundamentally "authoritarian scientific research" [37]. Their concerns are warranted, given the common use of field experiments without community consultation or consent [20]. In Campbell's view, participation and consent are basic requirements of an experimenting society.

Action researchers and qualitative researchers have also offered powerful critiques of the search for generalizable social science. Such research supports authoritarian governance when powerful people justify evidence-free policies using findings from another context [58, 37]. In an experimenting society, communities that consider a policy can develop their own evaluations. They also gain statistical benefits from adding their own situated experiments to a pool of common knowledge [10]. In Campbell's time, situated replication of policy evaluation was rejected as impossibly expensive [58].

With CivilServant, we are adapting the idea of an experimenting society to platform governance. By designing a system that supports plentiful, community-led policy evaluation, we are working toward an open society, where the public gains the benefits of experimental knowledge together with the benefits of a consequential voice on delegated platform governance. This paper 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 [65, 7]. When for-profit companies were permitted to operate online in the 1990s, volunteer *community leaders* on AOL governed its many chatrooms [64]. Internet users continue to create and enact policy on many major internet platforms, including Wikipedia [28], Facebook [26], Twitter [54, 30], reddit [51], and Xbox [34]. Responsibilities for identifying and responding to copyright violations and child pornography are delegated to third-party corporations [70] and nonprofits [72], an approach that some legal theorists have suggested for hate speech [16]. The accountability of these delegated authorities ranges widely, from communities with elections to organizations that operate in secret.

In the first deployment of CivilServant, we 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 associated with platform Terms of Service and computer fraud regulations [68]. On reddit however, independent data collection is routine to community moderation [45]. A strike by over a thousand reddit communities in 2015 demonstrated moderators' appetite for participatory policy-making [12]. This strike also revealed ways that moderators manage community expectations in their uses of power [52]. Based on this research, we chose reddit as our site because its communities already form a disputatious network of delegated governance power.

EXPERIMENTATION INFRASTRUCTURES

Systems supporting randomized trials are now a common component of social technologies, and software engineers and designers work in a process of "continuous experimentation" that in some firms evaluates tens of thousands of design interventions per year [44, 46, 78]. While the design of these systems varies between firms, platform-centered experimentation infrastructures tend to have common goals. These include making field experiments an efficient form of everyday software quality testing and making high quality field experiments accessible to software engineers and designers.

Creators of corporate experimentation infrastructures often describe their systems as an effort toward a "culture of experimentation" across a company [62]. By testing design changes that range from color decisions to major feature offerings, companies can learn their effects on user behavior, advertising outcomes, sales, and other company metrics. Companies attribute millions of dollars in increased revenue to evidence from field experiments [78] and report substantial cost savings in the allocation of work to designers and engineers [62]. In a culture of experimentation, these A/B tests are a basic part of the product development process across a firm.

Teams that manage experimentation infrastructures work to achieve a culture of behavioral testing with evolving pipelines of features that (a) monitor user behavior across a product, (b) coordinate interventions (c) and support the design and interpretation of research by non-experts. Such systems include Microsoft's EXP platform, LinkedIn's XLNT infrastructure, and AirBnb's ERF framework [46, 78, 61]. Fabijan describes the development of experimentation systems as an evolution from an early "crawl" stage with small numbers of bespoke experiments, to a "walk" stage of reproducible code, through stages of growing scale and adoption to a point where product teams can "fly." In this latter scale, much of the work of designing, monitoring, halting, analyzing, and making decisions based on experiments is automated [25]. For example, Facebook's PlanOut is a walk-stage system without a user interface that offers a engineers a domain-specific language for integrating randomized trials into their code [3]. The fly-stage system Wasabi supports designers to develop and deploy experiments without any knowledge of code [48].

While experimentation systems have broadened access to the means of experimentation within platforms, streamlined experiments depend on keeping participants uninvolved and unaware of research. None of these systems have publicly-documented features for informing or debriefing users; deception-based studies are the default. While deception is justifiable in social research, thousands of studies per year can enable decision-making that accumulates into abuses of behavioral knowledge. For example, on the ride-hailing platform Uber, the extensive use of deception-based studies has created what Rosenblatt and Stark call an "information asymmetry" that they argue can enable abuses of platform power [67]. This behavioral research has allegedly been used to manipulate drivers to benefit the company against drivers' own economic interests without their knowledge [69].

While experimentation systems can deliver mutual benefits to platforms and society, currently-implemented systems advance Popper's closed society. They foreclose awareness, critique, and the rejection of harmful interventions by the people whose behaviors they manage. CivilServant is our attempt to re-imagine the design of these systems for greater leadership by affected communities. In this paper, we present results from the bespoke, walk-stage of our system in hopes of developing higher volumes of publicly-accountable behavioral research in an open, experimenting society.

SUCCESSOR SYSTEMS

Systems like CivilServant collect data and coordinate users within a larger platform, creating alternative knowledge and supporting community mutual aid. Creators of these "successor systems" often attempt to restructure a larger system's power relations through software, data, and collective action, according to Geiger [29]. For example, with Turkopticon, digital laborers share mutual-aid information to evaluate the people who offer them work [41]. Community moderation technologies serve similar functions by developing independent knowledge and supporting community processes to restructure public life [30, 54]. Since social change is often a primary goal of these systems, they often function as critical infrastructure, generating ongoing knowledge that serves community needs [40]. With CivilServant, we applied this mutual aid, critical infrastructure approach to restructuring the power of experimentation.

DESIGN CONSIDERATIONS

Community Participation

Any process for evaluating social interventions will structure stakeholder power in some way. Some participatory evaluators prioritize close collaboration with existing power structures while others prioritize direct work with those who hold the least power [19]. Platform governance brings together a complex network of actors across communities, preventing power from being so easily classified. For example, moderation often involves exercising power in multi-party conflict situations in communities that may include tens of millions of people [43]. While platforms delegate more power to moderators than many other users, those moderators operate within systems defined by platforms and overlapping legal regimes. Furthermore, some of the least empowered people in platform governance are those who allegedly organize to harm

others. Because governance research often focuses on risk and harm, no research process can protect the most vulnerable while guaranteeing equal participation of all stakeholders.

We have designed CivilServant to expand the potential participation of anyone involved in or affected by online moderation. We draw inspiration from Arnstein's ladder of citizen participation, which poses a scale from non-participation to tokenism to citizen power [2]. While we cannot and should not offer equal power to all stakeholders, we can offer some participation to everyone involved. By supporting communities to lead their own studies, publishing all findings, and openly discussing results, we expand participation for all stakeholders, including moderators, casual contributors, and those who are judged to violate community policies.

Research Ethics

Social computing researchers are currently re-negotiating research ethics after highly-publicized controversies over corporate and university experiments conducted without consent or ethics board review [35, 27]. In the U.S., a mismatch between medical research ethics and social research has created an urgent need for progress on research ethics [74].

Given the tremendous power exercised by those who govern online behavior, we take the view that power-holders have an ethical obligation to evaluate governance outcomes [56]. We also believe practical, participant-led evaluations of attempts at social change can contribute to social scientific knowledge [60] while expanding options for consent in social research.

Large-scale policy experiments necessarily entail complex relations of risk and harm. Because our research focuses on governance, we have drawn inspiration from recent conversations about field experiments in political science, where multi-party interests and public goods often conflict in complex ways [21]. To manage these risks, political scientists are developing novel methods for consent from groups, stakeholders, and participant representatives, as well as novel debriefing procedures [20]. One design goal for CivilServant is to support empirical research on novel ethics procedures; in this paper, we report results from community debriefings.

Open Knowledge and Transparency

We created CivilServant to generate open knowledge. Since the primary audiences for that open knowledge are community members who may just be starting to develop their data literacy, we prioritize the general-audience publishing and community engagement needed to reach communities with our findings. By publishing all software and analyses, we are also able to contribute to an open research culture among other scholars who can query and replicate our findings [57].

While designing CivilServant, we also encountered an apparent tension between research openness and privacy that we have yet to resolve. With open datasets, scholars and participants can confirm, question, and extend our findings by reanalyzing research data. Since these datasets also represent privacy risks, we currently keep all experiment data private. We may consider releasing data of future studies if we can develop community-accepted processes for data publication.

Deliberative Replication

In Campbell's imagined experimenting society, randomized trials are a plentiful form of knowledge generated by citizens who develop their own situated knowledge rather than rely on studies conducted elsewhere [10]. Experimentation infrastructures advance this goal by expanding the work of experiments to non-experts and by facilitating the re-use of measurements and study designs. In the design of CivilServant, we use similar means to support community replications that prioritize each community's goals. Consequently, where designers of other experimentation systems often value automated decisionmaking based on evidence from replicated findings, we give priority to community deliberation.

THE CIVILSERVANT SYSTEM

Like other similar infrastructures, CivilServant monitors user behavior, manages sampling and treatment, generates variables, and supports analysis. Designers of systems at this stage of mostly-bespoke research tend to describe on the research process and reusable components [25]. Since community-led policy experiments differ substantially from platform experiments, we first report the kinds of research our system supports, then the experiment process, and finally the system features that carry out this process.

Supported Experiment Designs

While not all experimentation infrastructures have an end-user interface [3], they all do some work to monitor behavior, coordinate interventions, and support the design and interpretation of research. Like other early-stage systems, CivilServant has no graphical user interface for research design or management. Instead, researchers utilize re-usable system features and software modules to manage routine parts of the research process. Replications that use these modular features can be configured through a domain-specific language, while novel procedures require bespoke software.

The system presents itself to a community in the form of a bot account [30], which moderators grant permission to be a moderator. As a moderator account, the system is prominently displayed to users in the subreddit page, and all of its public actions are visible to anyone on reddit. The levels of permissions granted to this account by moderators determines the data collection it is capable of in a given context (Table 1).

While some experiments involve direct interventions to users or discussions from the bot account, CivilServant can also alter community-wide configurations in time-randomized studies or induce humans to carry out interventions at individual and group levels. CivilServant currently supports research designs that test many of the interventions available to volunteer moderators on reddit (Table 1).

CivilServant supports multi-armed, conditional assignment on user accounts and discussions, including stratification, block randomization, and cluster assignment [32]. Real-time enrollment can be done with automated event detection or community labeling. The system also supports assignment over time, where community-wide interventions such as interface designs are shown during randomly-selected periods.

Variables	Interventions	Assignment
Counts of comments	**Public**	Discussions
Counts of posts	Announcements	Comments
Comment/post removal	Messages	Users
Newcomer counts	Hide interfaces	Time Periods
Vote scores	Show interfaces	Occurences
Post rankings	**Private**	
Post scores	Suspend users	(assignments)
User political leaning	Ban users	(can be)
Recidivism rates	Private messages	(conditional)
User experience	Inducements	

Table 1. Some re-usable experiment elements supported by CivilServant

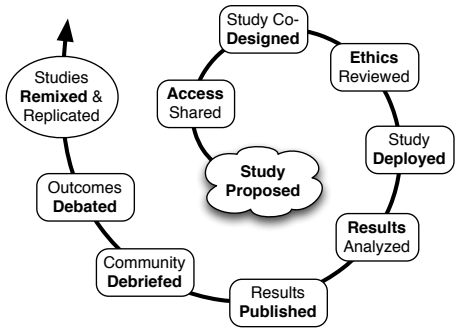

Figure 1. The Community Knowledge Spiral: CivilServant studies use processes that grow community-led experimentation.

As a system that works directly with users, connects to platform APIs, and operates independently from privileged platform access, CivilServant can only observe data routinely shared by a platform with its users. Since reddit users are also limited in these ways, these constraints have not limited our ability to evaluate questions asked by reddit users.

System Architecture

We designed CivilServant to aid study design, collect data, manage interventions, and support our processes for analysis, reporting, and debriefing. Created in Python, R, and MySQL, the system was managing hundreds of millions of records across a distributed infrastructure at the time of writing.

We designed the system to enable expansion to multiple platforms. A domain-specific language similar to PlanOut [3] describes the platform, the 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 communities. At the time of writing, CivilServant connects to reddit, Twitter, and third-party systems used by moderators.

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 generate experiment results and reports using R Markdown, as specified in our pre-analysis plans. Our repositories are linked with the Open Science Framework, which hosts all pre-analysis plans [57].

Designing Studies with CivilServant

Studies conducted with CivilServant follow a social and technical process that we call the community knowledge spiral (Figure 1). While only researchers directly control the system, details of a study are developed by the community. 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. Because reddit communities already collect and process data, many have strong intuitions about interventions and measures that would be feasible in a field experiment. CivilServant researchers convert these ideas into study designs by facilitating discussions on community goals that the software can support.

Permission for Data Access

If moderators wish to continue, they invite the CivilServant reddit account to become a moderator with archival-only privileges to their community. At this point, the system can collect historical and ongoing data on submitted posts, comments, ranking algorithm behavior, moderator actions, and data from other bespoke systems that moderators use to coordinate their work. As an account with archival-level access, CivilServant can process data and support further planning while the community is still discussing whether to trust the system to intervene. By connecting CivilServant to communities through a user account, we establish an important basis for group consent: communities can observe the presence of the CivilServant system and revoke its data access any time.

Study Design, Power Analyses, and Pre-Analysis Plans

After CivilServant has collected data for a period of time, we use the data to construct a formal study design from the community requests. A narrative pre-analysis plan provides the community with a formal description in non-expert language that explains a study's goals, interventions, variables, and analysis procedures [47]. 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, power analyses inform the community about their chance of observing certain effect sizes in the time-spans they are interested to conduct the study.

University Ethics Approval

Once communities decide the area of policy and the kinds of measures used in a study, we 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. 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 individual consent but requires communities to be debriefed, also excludes studies involving banning accounts and other interventions carrying social costs that are not easily reversible. Most routine policy evaluations conducted by CivilServant fit within this IRB. We have also sought and received per-study IRB approval when the risks are higher, or when individual consent and debriefing are more appropriate.

Finalizing Study Designs

When moderators and communities discuss a pre-analysis plan, they may notice details needing adjustment or changes to their own moderation infrastructure. For example, participants may think of potentially-confounding factors on the platform or within their community, factors that we add as covariates for regression adjustment. Researchers configure the final study design with a domain-specific language.

Recipes for Theory-Informed Interventions

CivilServant study designs sometime test phrasing variations in announcements and personal communications. When communities design a study, we support them to adopt theoretically-informed language by listing a "recipe" of guiding social theories. For example, in statements of rules, a community might wish to appeal to widely-held norms, to authority, or to enforcement consequences. Using a collaborative text editor, we provide a list of possible "ingredients" from social theory and suggest them as starting-points for messaging alternatives that they choose as a group.

Testing & Deploying Community Experiments

Once a community's moderators agree to a final experiment design, we publish the final pre-analysis plan and test the experiment software for compatibility with the community's other systems. Next, we generate reproducible randomizations and deploy the experiment for the agreed-on period, monitoring experiment activity for compliance to the study procedures. All studies are block-randomized; when software or compliance errors occur, individual blocks can be removed without spoiling the balance of the sample.

Concluding Community Experiments

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

Analysis of Findings

At the conclusion of a study, the CivilServant software generates dataframes for each of the hypotheses in the study. We then produce an experiment report that applies statistical methods listed in the pre-analysis plan to the observed data. For study replications, this process is automated.

Debrief and Discuss With Community

We prioritize maximum disclosure to participating communities. Where possible, we host public discussion of our research with a community during the study design process. We make all study results public. We also require that all participating communities agree to host a "community debriefing," a public conversation within their community to report and discuss the results. To support that conversation, we publish a public-audience summary of the study motivations, procedures, and findings on the CivilServant website. We 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. We 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

We chose to deploy CivilServant first 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. As with new studies, researchers manage the replications design process. These deliberative community replications are central to our goals of fostering an experimenting society online. 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.

EVALUATING CIVILSERVANT

To evaluate CivilServant in this early stage, we take approaches from research on system design, critical infrastructure, and policy evaluation. In computer science, systems papers about early-stage experimentation infrastructures tend to report design considerations, describe implementation details, summarize early experiments, and discuss the role of experiments in software engineering organizations [46, 3, 78, 25]. Critical infrastructure systems such as Turkopticon have typically been evaluated for their role to foster reflection and collective organizing among those who use or encounter it [29, 41, 22, 40]. In policy fields, researchers have struggled to compare the policy contributions of different experimenting approaches because policy evaluation is a complex political process that often begins and ends with group decisions [18]. Instead, the policy fields evaluate an experimentation approach by its methodological validity, by its adherence to stated values, and by the uses of research findings [19, 18].

Similarly to other early-stage research on experimentation infrastructures, we report two case studies of large-scale community conducted by reddit communities. Detailed experiment findings are being published elsewhere; we present them here as cases in the design of the CivilServant system and context for our qualitative findings. Because we designed CivilServant as critical infrastructure, we also report qualitative findings on the kinds of critical perspectives that over a thousand participants brought to community debriefings.[1] In these qualitative findings, we evaluate CivilServant as a process for community-led policy evaluation, and reflect on what we as designers and researchers learned about the possibility of a democratic experimenting society online.

CivilServant Case Studies

To illustrate the kind of research that CivilServant supports, we offer two case studies of community research.

[1] We have obfuscated all quotations from these debriefings.

Increasing Newcomer Norm Compliance

Moderators of r/science, a 13-million subscriber community at the time, approached us to conduct the first CivilServant study, which we supported from August 25 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 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.

Moderators used CivilServant to evaluate the effect of posting announcements of community rules to the top of discussions. The CivilServant bot randomly assigned announcements to some discussions and not to others. The system observed discussion removals and comment removals, blinding moderators to the treatment condition. CivilServant also observed the number of minutes that discussions were promoted by reddit's algorithms. Interventions were block-randomized based on the type of discussion. Across 2,190 discussions of academic publications and 24 live question-answer sessions, we found that without posting policy announcements, a first-time commenter has a 75.2% chance of complying with community rules, and that posting the rules has a positive 7.3 percentage point effect on the chance that a newcomer's first comment will be allowed to remain by moderators, on average in the community. Although we expected that posting the rules more visibly would reduce newcomer participation, 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 unruly behavior each month [redacted].[2]

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 reddit communities also discussed the findings. 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 year, the reddit company enrolled nearly a hundred communities in a replication of this study. Employees at platforms beyond reddit took notice of this study. In August 2017, Disqus credited the study by r/science as a motivation for new features they provided to over a million websites [39].

Moderators and community commenters contributed substantial ideas to the design and interpretation of this study, many of which we incorporated into CivilServant for future studies:

Moderators designed ways to be blinded during the study. During study design discussions, one moderator suggested that since the outcome variable relied on comment removals, moderators should be blinded to prevent bias from knowledge of which treatment arm had been assigned. They prototyped and deployed stylesheet adjustments which would hide all treatments from moderators only. This stylesheet is now standard for all announcement-based reddit experiments.

Moderators proposed variables and methods for block randomization and regression adjustment. As moderators discussed the study design, some argued that Q&A discussions needed different kinds of moderation. They suggested that our software could automatically identify the conversation type based on "flair" labels from moderators. We have reused this block-randomization approach in later studies.

Some moderators wondered if experiment validity might be affected if reddit's algorithms promoted some discussions to greater numbers of newcomers. We extended CivilServant to monitor reddit's rankings over time to adjust for this factor. This ranking monitor was unexpectedly useful in our second study when we used it to construct the main outcome variable.

Managing Verification & Promotion of Unreliable News

Moderators of the r/worldnews community approached us in October 2016 to test methods for governing the reception and spread of news from frequently-inaccurate sources in their subreddit. Articles from these web domains were 2.3% of all submissions to the community, which reviewed an average of 450 articles per day. Moderators wished to avoid banning often-erroneous websites but also wished to encourage reader skepticism toward them. We were also concerned that if our intervention increased fact-checking activity, this behavior might influence reddit's news recommenders, causing fact-checked articles to be promoted in the rankings.

In this multi-armed study design, our software posted announcements encouraging readers to fact-check articles by linking to further evidence. In a second message, moderators added further language encouraging readers to use reddit's voting systems to dampen the algorithmic spread of these 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. In a forthcoming paper, we show that all arms that increased the chance that individual comments and discussions would include links to further evidence [redacted].[3] However, while the arm encouraging fact-checking caused the algorithmic ranking of news articles to be demoted by as many as 24 rank positions over seven hours in the top 300 entries on average, we failed to discern an effect on news rankings from the intervention that also encouraged voting. The findings confirmed our expectation that encouraging fact-checking could influence reddit's rankings, but the outcome was the opposite of 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. Other reddit communities also discussed the results. We also received over a dozen personal notes from r/worldnews participants about the study. By the summer, moderators voted to adopt the encouragement that

[2] reference omitted for review

[3] reference omitted for review

readers fact-check articles, crowdsourced a list of sites from readers, and developed their own bot to carry out the policy.

Moderators noticed platform changes affecting the study. During this study, the reddit platform altered the function of their news aggregation algorithms. Moderators noticed the change while the study was active and alerted us. Their close attention to everyday experience allowed us to identify invalid assignments and extend the study while it was still underway.

Studies In Progress

We have worked with communities on a wide range of evaluations beyond these case studies. For example, further studies include efforts at conflict resolution in polarized discussion groups, peer responses to identity-based attacks, and interventions to mitigate the social side effects of unreliable machine learning moderation. Two subreddits have decided to test policies that may reduce recidivism rates among participants who are re-integrated into their communities after being banned. Another four communities are considering replications of the experiment first conducted by r/science.

DELIBERATION AND USES OF CIVILSERVANT FINDINGS

Because CivilServant is designed to advance the values of an open society where people affected by research can deliberate the implications, we also offer early impressions from discussions in the two case studies we have presented.

Findings on community debriefings are based on participant observation, conversation logs, and associated field notes. Communities held these debriefings by posting the 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. One of the researchers answered community questions during debriefings.

Findings on the usage of study results are informed by interviews with over a dozen moderators, participant observation in text conversations with over a dozen subreddits, field notes from the experiment design process, and emails exchanged with 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 by linking to results, or if their moderators responded to a recruitment message posted to moderator discussion areas. Moderators were sampled for interview from communities that discussed and adopted and those that rejected the evaluated.

Discussion of Findings in Community Debriefings

In debriefings, *many participants shared personal stories from the experiment*. One person in the news study 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." These stories often opened longer discussions about the purpose or legitimacy of moderation policies. One commenter reflected that "After I start typing, I see that a rule that conflicts with my comment and curse." When someone replied "Isn't that the point?" commenters discussed whether the outcome was beneficial or not.

Participants sometimes offered direct critiques of community policies. For example, some argued that the science discussion community should permit contributions from climate change skeptics. In the news community, some commenters argued that moderators should have included state-sponsored media in the fact-checking intervention. When one commenter complained that encouragements to fact-check amount to telling readers how to think, other commenters argued that the intervention encouraged critical thinking and greater intellectual independence among readers—outcomes that our research had not directly observed.

Commenters shared questions and critiques of research methodology. 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. We were surprised by the number of people with knowledge of research methods in both communities; *many statistics questions were answered other community members*. The demographics of reddit may explain why community members answered many of the statistical questions. On average in the U.S., 82% of reddit users have some college education, twenty-three percentage points more than the rest of the population. The difference is even higher among reddit users who browse the site for news [5].

Commenters offered personal theories to explain experiment results. For example, many questioned whether effects would endure over time. Others described 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 books such as Kahneman's *Thinking Fast and Slow*.

In both debriefings, *participants discussed whether the evaluated policy might be useful in their other communities*. Roughly 15% of the comments in one of the debriefings 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 policy elsewhere. Some argued that we should have withheld sharing any results until completing further replications.

Commenters in both debriefings discussed research ethics. Several community members argued that we should release full datasets, leading to extended discussions of our 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 [35]. In the discussions that followed, arguments over research ethics were interleaved with arguments over community policies. One participant argued that since community policies against abusive speech and personal attacks were an unjust form of censorship, experimental interventions that reduce the rate of abusive speech are unethical. Elsewhere, one debriefing included an extended discussion about the justice of community policies and U.S. research ethics regulations.

Many people expressed gratitude for research findings in community debriefings. Having prepared for possible responses similar to public outrage over controversial platform research, we were surprised at the volume of appreciation, in private and public message. Many of these 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 we shared results from our fact-checking study fewer than two weeks after the 2017 US presidential inauguration, we expected that some US commenters would interpret our work as politically-partisan. Instead, people of all political affiliations thanked us and the 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 community criticism. They expected debriefings to attract complaints. Others disagreed. One moderator 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 moderators expected the findings would be popular, since readers frequently complained about inaccurate stories and often reply to articles with profanity-filled complaints. This moderator expected that participants would the experiment as an effort to respond to community demands.

Community Uses of Experiment Findings

While the first findings from CivilServant were only published less than a year ago, our qualitative research on the uses of CivilServant results provides an early perspective on the uses of knowledge from community-led experiments.

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 [18]. 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 than as a judgment on the effectiveness of a specific intervention [76]. Yet research read for general enlightenment can, in time, inform those external forces as well [77]. In moderator interviews, content analysis of subreddit discussions, and correspondence with reddit employees, we found that communities' uses of CivilServant findings follow many of these usage patterns and constraints.

Community policy adoption can take months and may not be predicted by beneficial results. While moderators can apply many policies instantly by reconfiguring their software, decisions occur more slowly. For example, moderators of one community first discussed policy changes six months after completing the study. Within a few days, they decided to adopt the policy. In an interview three months before the decision, a moderator explained that they hoped to make changes, but more pressing demands had prevented them from finding time to reconfigure their automated moderation system. Nuances of deployment prevented another community from implementing a policy they agreed was beneficial. In interviews, moderators described the difficulty of reaching agreement about the wording of the policy, and discussions stalled.

How Other Communities Used Results

We designed CivilServant to generate knowledge that could be shared, debated, and replicated by a growing network of communities. Here we report early findings from discussions beyond the communities in our case studies.

Moderators used research findings to advocate for change, develop replications, and defend existing policies. In interviews, several reported suggesting that their community adopt the policies we tested. In one case, a community justified automated policy announcements by linked directly to our study results. Elsewhere, when some expressed skepticism that research findings would apply to their subreddits, the moderation team contacted us to conduct replications. Participants in another subreddit appealed to our findings to defend similar policies that predated our research. Regular participants had complained that visible listings of community policies were annoying and ineffective. In interviews, moderators reported that they discussed our findings in the discussion where they chose to retain the practice.

Research also informed moderators' personal practices when group-wide policies were implausible. For example, when moderators of one gaming community with over half a million subscribers read the r/science experiment results, they considered automating announcements with the rules. After a group decision on an automated system did not materialize, individual moderators decided to personally-post announcements 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. While these trials were not randomized, they represent efforts to develop situated knowledge based on community experiment findings.

How Platforms Used Results

While we designed CivilServant to create platform-independent research, community interests often align with the interests of platforms. In the case studies reported here, we notified the company about our findings after debriefing the community. 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 company had enrolled nearly a hundred communities in a voluntary randomized trial replicating our results. [4] Our results were also cited by Disqus, who credited the study on increasing new norm compliance when announcing new features and advising moderators how to use them [39]. In both cases, our findings informed platform features and research that foregrounded community leadership on policy.

DISCUSSION

Our work with CivilServant addresses a complex dilemma for governance in the 21st century. Because platforms monitor and intervene in many people's lives, governance initiatives that work through platforms might advance justice, improve

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

well-being, expand understanding, and save lives. Experimentation infrastructures could substantially advance society's ability to develop and choose effective policies, but current approaches to designing secret, large-scale research also represents a very serious risk to open, democratic societies.

In this paper explore the idea of a democratic, experimenting society online by introducing CivilServant, a novel system for community-led experimentation. Building on a thought experiment by policy evaluation pioneer Donald Campbell and on traditions of action research, we have prototyped a system that allows communities with delegated governance power to conduct their own, publicly-accountable field experiments, independently from oversight by platforms. We have situated CivilServant in relation to other experimentation infrastructures, reported on our design considerations for the software and research processes, described the system design, shared two case studies, and discussed early community responses and uses of experiment knowledge.

Using CivilServant, communities on reddit have evaluated practical ideas for preventing thousands of comments that would otherwise be removed by moderators and managing the promotion of news from frequently-inaccurate sources. In both of the reported case studies, experimental evidence contradicted the best predictions of community moderators and the researchers. Public debates about experiment results gave communities a chance to decide against ineffective policies and apply governance techniques that achieve their goals in their own community on average.

Across these case studies, we find that our participatory process improved the quality of experiment designs. Planning discussions with communities improved the design of experiment outcomes, adjustment variables, assignment, intervention procedures, and estimation strategies. Experimenters are sometimes imagined as autocrats who force procedures onto participants in pursuit of validity. We have found that communities sometimes request personal inconveniences in the interest of validity when they lead the research goals—illustrated by moderators who proposed that they be blinded in their own study.

As a systems paper, our experience with CivilServant can be seen as a kind of replication of findings on early stage experimentation infrastructures. When other researchers find from case studies that "controlled experiments can be run and their results are trustworthy" [46], they imply complex arguments about the mechanisms of research and the power of stakeholders whose trust the system is designed to cultivate. While CivilServant uniquely includes community moderators and research participants in that circle of accountability and trust, its evolution has paralleled other systems. Fabijan's model for theorizing the evolution of such systems has predicted many of our project's technical needs as it has grown. For example, as we moved from bespoke experiments to a "walk" stage of reusable software models, we found ourselves developing pre-computed variables, common intervention modules, and systems supporting power analyses (Table 1) [25].

We designed CivilServant hoping that experiments could contribute situated knowledge to an open society. Personal stories about experiment results, community replications, and informal personal trials of policy ideas all illustrate community approaches to situated experimentation. The clearest example occurred when communities re-appropriated a variable monitoring aggregator rankings. While the science community used the variable for regression adjustment, the worldnews community put this measure at the heart of a completely new study focusing on algorithm behavior.

Community-led experiments also attracted substantive deliberation on policy decisions and the research process itself. In community debriefings, participants critiqued policies, questioned findings, and developed theories to explain the results. While some might see this deliberation as a delay in the adoption of evaluated policy ideas, we take these early results as evidence that experimenting communities can hold open and nuanced conversations about the relative merits of experimental evidence when governing their own affairs.

Like other participatory policy work, the CivilServant software and research process cannot offer equal power to every participant. Without the ability to observe individual viewing behavior, we are sometimes unable to observe how many people received a treatment or what proportion of participants were aware of or represented in community debriefings. Further work on the politics and ethics of community-led platform experiments will need to develop principles and mechanisms for assessing outreach and participation in deliberation and accountability processes.

In our case studies, community debates about research ethics add further evidence to the argument that ethics discussions often concern the construction of power in society and the ways that high status people use their power [6]. Community-led experiments offer alternative power relations from platform experiments, but community leaders in participatory processes are also capable of abuse. While public debriefings do enable some degree of moderator accountability, further work is needed on principled approaches to the ethics and politics of community experiments.

Can community-led experiments ever reach the scale required to meaningfully-advise the use of platforms to govern society in an open, democratic society? As platforms become a common point of intervention for governing societal risks, large-scale behavioral research could make policymakers more effective and less accountable to the people they govern. With CivilServant, we have demonstrated that it is possible to redesign experimentation infrastructures for an open society. Given the implications for human flourishing and freedom, further progress on the politics and design of online experiments is urgently needed.

AUTHOR CONTRIBUTIONS STATEMENT

ACKNOWLEDGMENTS

Acknowledgments omitted for review.

REFERENCES

1. Adelman, C. Kurt Lewin and the origins of action research. *Educational action research 1*, 1 (1993), 7–24.
2. Arnstein, S. R. A ladder of citizen participation. *Journal of the American Institute of planners 35*, 4 (1969), 216–224.
3. Bakshy, E., Eckles, D., and Bernstein, M. S. Designing and Deploying Online Field Experiments. In *Proceedings of the 23rd International Conference on World Wide Web*, WWW '14, ACM (New York, NY, USA, 2014), 283–292.
4. Banet-Weiser, S., and Miltner, K. M. # MasculinitySoFragile: culture, structure, and networked misogyny. *Feminist Media Studies 16*, 1 (2016), 171–174.
5. Barthel, M., Stocking, G., Holcomb, J., and Mitchell, A. Reddit news users more likely to be male, young and digital in their news preferences. Tech. rep., Pew Research Center, Feb. 2016.
6. boyd, d. Untangling research and practice: What Facebooks emotional contagion study teaches us. *Research Ethics 12*, 1 (2016), 4–13.
7. Bruckman, A., Curtis, P., Figallo, C., and Laurel, B. Approaches to managing deviant behavior in virtual communities. In *CHI Conference Companion* (1994), 183–184.
8. Buni, C. The secret rules of the internet, Apr. 2016.
9. Campbell, D. T. Comment: Another perspective on a scholarly career. *Scientific inquiry and the social sciences* (1981), 453–501.
10. Campbell, D. T. The experimenting society. In *The experimenting society: Essays in honor of Donald T. Campbell*. Transaction Publishers, New Brunswick, 1998, 35.
11. Campbell, D. T., and Stanley, J. C. *Experimental and Quasi-Experimental Designs for Research*, 1 ed. Wadsworth Publishing, 1963.
12. Centivany, A., and Glushko, B. Popcorn Tastes Good: Participatory Policymaking and Reddit's. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, ACM (2016), 1126–1137.
13. Chan, M. ADL Tackles Hate Speech With Silicon Valley Command Center | Time.com. *Time Magazine* (Mar. 2017).
14. Chancellor, S., Pater, J. A., Clear, T., Gilbert, E., and De Choudhury, M. # 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*, ACM (2016), 1201–1213.
15. Cheng, J., Danescu-Niculescu-Mizil, C., and Leskovec, J. How Community Feedback Shapes User Behavior. *ICWSM 2014* (May 2014). arXiv: 1405.1429.
16. Citron, D., and Wittes, B. Follow Buddies and Block Buddies: A Simple Proposal to Improve Civility, Control, and Privacy on Twitter, Jan. 2017.
17. Citron, D. K., and Norton, H. L. Intermediaries and hate speech: Fostering digital citizenship for our information age. *Boston University Law Review 91* (2011), 1435.
18. Contandriopoulos, D., Lemire, M., Denis, J.-L., and Tremblay, . Knowledge Exchange Processes in Organizations and Policy Arenas: A Narrative Systematic Review of the Literature. *Milbank Quarterly 88*, 4 (Dec. 2010), 444–483.
19. Cousins, J. B., and Whitmore, E. Framing participatory evaluation. *New directions for evaluation 1998*, 80 (1998), 5–23.
20. Desposato, S. Ethical Challenges and Some Solutions for Field Experiments.
21. Desposato, S. *Ethics and Experiments: Problems and Solutions for Social Scientists and Policy Professionals*. Routledge, 2015.
22. Dimond, J. P., Dye, M., Larose, D., and Bruckman, A. S. Hollaback!: the role of storytelling online in a social movement organization. In *Proceedings of the 2013 conference on Computer supported cooperative work*, ACM (2013), 477–490.
23. Doleac, J. L., and Stein, L. C. The visible hand: Race and online market outcomes. *The Economic Journal 123*, 572 (2013), F469–F492.
24. Edelman, B. G., Luca, M., and Svirsky, D. Racial discrimination in the sharing economy: Evidence from a field experiment.
25. Fabijan, A., Dmitriev, P., Olsson, H. H., and Bosch, J. The Evolution of Continuous Experimentation in Software Product Development. In *International Conference on Software Engineering (ICSE)* (2017).
26. Facebook. Group Admin Basics: What is a Group Admin?
27. Fiesler, C., Chancellor, S., Hoffmann, A. L., Pater, J., and Proferes, N. J. Challenges and Futures for Ethical Social Media Research. In *AAAI Conference on Web and Social Media (ICWSM): Workshop* (2016).
28. Forte, A., Larco, V., and Bruckman, A. Decentralization in Wikipedia governance. *Journal of Management Information Systems 26*, 1 (2009), 49–72.
29. Geiger, R. S. Successor Systems: The Role Of Reflexive Algorithms In Enacting Ideological Critique. *Selected Papers of Internet Research 4* (2014).
30. Geiger, R. S. Bot-based collective blocklists in Twitter: the counterpublic moderation of harassment in a networked public space. *Information, Communication & Society 19*, 6 (June 2016), 787–803.

31. Geiger, S. Does facebook have civil servants? On governmentality and computational social science. In *Workshop on Ethics for Studying Sociotechnical Systems in a Big Data World* (Vancouver, British Columbia, Canada, 2015).
32. Gerber, A. S., and Green, D. P. *Field experiments: Design, analysis, and interpretation*. WW Norton, 2012.
33. Gillespie, T. The politics of platforms. *New Media & Society 12*, 3 (2010), 347–364.
34. Good, O. Does Your Gamertag Have Herpes? Beware Xbox Live Enforcement United. *Kotaku* (Aug. 2013).
35. Grimmelmann, J. The law and ethics of experiments on social media users.
36. Halfaker, A., Geiger, R. S., Morgan, J. T., and Riedl, J. The rise and decline of an open collaboration system: How Wikipedias reaction to popularity is causing its decline. *American Behavioral Scientist* (2012), 0002764212469365.
37. Hayes, G. R. The relationship of action research to human-computer interaction. *ACM Transactions on Computer-Human Interaction (TOCHI) 18*, 3 (2011), 15.
38. Hill, B. M., and Shaw, A. Studying Populations of Online Communities. In *The Handbook of Networked Communication*. Oxford University Press, New York, NY, 2017.
39. Hue, T. How to create an effective Comment Policy that readers actually follow, Aug. 2017.
40. Irani, L., and Silberman, M. From critical design to critical infrastructure: Lessons from Turkopticon. *interactions 21*, 4 (2014), 32–35.
41. Irani, L. C., and Silberman, M. Turkopticon: Interrupting worker invisibility in amazon mechanical turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, ACM (2013), 611–620.
42. Johnson, L. B. 296 - Memorandum on the Use and Management of Computers by Federal Agencies, June 1966.
43. Keegan, B. C., and Matias, J. N. Actually, It's About Ethics in Computational Social Science: A Multi-party Risk-Benefit Framework for Online Community Research. *arXiv preprint arXiv:1511.06578* (2015).
44. Kevic, K., Murphy, B., Williams, L., and Beckmann, J. Characterizing experimentation in continuous deployment: a case study on bing. In *Proceedings of the 39th International Conference on Software Engineering: Software Engineering in Practice Track*, IEEE Press (2017), 123–132.
45. Kiene, C., Monroy-Hernndez, A., and Hill, B. M. 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*, CHI '16, ACM (New York, NY, USA, 2016), 1152–1156.
46. Kohavi, R., Deng, A., Frasca, B., Walker, T., Xu, Y., and Pohlmann, N. Online Controlled Experiments at Large Scale. In *Proceedings of the 19th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '13, ACM (New York, NY, USA, 2013), 1168–1176.
47. Lin, W., and Green, D. P. Standard operating procedures: A safety net for pre-analysis plans. *PS: Political Science & Politics 49*, 3 (2016), 495–500.
48. Lita, L. Meet Wasabi, an Open Source A/B Testing Platform, Jan. 2017.
49. Lomas, N. Twitter nixed 635k+ terrorism accounts between mid-2015 and end of 2016. *TechCrunch* (Mar. 2017).
50. MacKinnon, R. Consent of the networked: The worldwide struggle for Internet freedom. *Politique trangre 50*, 2 (2012).
51. Matias, J. N. The Civic Labor of Online Moderators (Oxford, UK, Sept. 2016).
52. Matias, J. N. 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*, ACM (2016), 1138–1151.
53. Matias, J. N. A toxic web: what the Victorians can teach us about online abuse. *The Guardian* (Apr. 2016).
54. Matias, J. N., Johnson, A., Boesel, W. E., Keegan, B., Friedman, J., and DeTar, C. Reporting, Reviewing, and Responding to Harassment on Twitter. *arXiv preprint arXiv:1505.03359* (2015).
55. Metz, R. Facebook Lives new suicide-prevention tools come with good intentions but many questions. *MIT Technology Review* (Mar. 2017).
56. Meyer, M. N. Two Cheers for Corporate Experimentation: The A/B Illusion and the Virtues of Data-Driven Innovation. SSRN Scholarly Paper ID 2605132, Social Science Research Network, Rochester, NY, May 2015.
57. Nosek, B. A., Alter, G., Banks, G. C., Borsboom, D., Bowman, S. D., Breckler, S. J., Buck, S., Chambers, C. D., Chin, G., Christensen, G., and others. Promoting an open research culture. *Science 348*, 6242 (2015), 1422–1425.
58. Oakley, A. Experiments in knowing: Gender and method in the social sciences.
59. O'Donovan, C. Nextdoor Rolls Out Product Fix It Hopes Will Stem Racial Profiling. *BuzzFeed* (Aug. 2016).
60. Paluck, E. L., and Cialdini, R. B. Field research methods. *Handbook of research methods in social and personality psychology* (2014), 81–97.

61. Parks, J. Scaling Airbnbs Experimentation Platform, May 2017.

62. Pettingill, L. M. 4 Principles for Making Experimentation Count, Mar. 2017.

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

64. Postigo, H. America Online volunteers. *International Journal of Cultural Studies 12*, 5 (Sept. 2009), 451–469.

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

66. Roberts, S. T. Commercial Content Moderation: Digital Laborers' Dirty Work.

67. Rosenblat, A., and Stark, L. Algorithmic Labor and Information Asymmetries: A Case Study of Ubers Drivers. SSRN Scholarly Paper ID 2686227, Social Science Research Network, Rochester, NY, July 2016.

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

69. Scheiber, N. How Uber Uses Psychological Tricks to Push Its Drivers Buttons. *The New York Times* (Apr. 2017).

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

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

72. Thakor, M., and others. Networked trafficking: reflections on technology and the anti-trafficking movement. *Dialectical Anthropology 37*, 2 (2013), 277–290.

73. Thakor, M. N. *Algorithmic detectives against child trafficking: data, entrapment, and the new global policing network*. PhD thesis, Massachusetts Institute of Technology, 2016.

74. Vitak, J., Shilton, K., and Ashktorab, Z. 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*, ACM (2016), 941–953.

75. Wark, L. Inside Alphabets Jigsaw, the powerful tech incubator that could reshape geopolitics. *Quartz* (Nov. 2016).

76. Weiss, C. H. Research for policy's sake: The enlightenment function of social research. *Policy analysis* (1977), 531–545.

77. Weiss, C. H. The many meanings of research utilization. *Public administration review 39*, 5 (1979), 426–431.

78. Xu, Y., Chen, N., Fernandez, A., Sinno, O., and Bhasin, A. From infrastructure to culture: A/b testing challenges in large scale social networks. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM (2015), 2227–2236.