RESEARCH STATEMENT

J. NATHAN MATIAS

How do individual behavior and social structures change each other and how do software algorithms influence these dynamics in online communications? To understand *behavior change in groups and networks*, I have conducted field experiments and observational studies to discover (a) how group norms influence online harassment and (b) how people change norms and structures of inequality. To study *inter-dependencies in human and machine behavior*, I have investigated how nudging human behavior also influences the decisions of algorithms to promote misinformation.

As I pursue these questions, I develop methods and software to remake large-scale behavioral research for democracy in the digital era. The nonprofit I founded, CivilServant, supports my research through *citizen behavioral science*. CivilServant works alongside the public to discover the outcomes of ideas for a flourishing internet and to test the social impacts of technologies in our digital lives. My interdisciplinary action research consists of field experiments, digital ethnography, and system design that contribute to communications, social psychology, and human-computer interaction.

BEHAVIOR CHANGE IN GROUPS AND NETWORKS

As the pioneering social psychologist Kurt Lewin argued in the 1940s, conflict and prejudice involve people acting within social structures, and attempts to resolve them require attention to group dynamics (Lewin, 1948). These enduring global problems now have digital dimensions, creating pragmatic needs to manage these problems. I am drawn to field research that develops and validates theory by testing attempts to change individual and group behavior.

Unruly, harassing behavior is common online, forcing many people to avoid participation in public discourse. Theories of human behavior suggest that people's decisions to participate in a group and their subsequent behavior are influenced by knowledge of what is socially normative. Visible community rules against harassment could reduce fears that prevent people from participating while also reducing harassment among those who do join. I tested these theories in a field experiment under review by randomizing announcements of community rules to over a thousand online conversations in a science discussion community with 13 million subscribers. Compared to no mention of rules, the announcements increased newcomer rule compliance by over 7 percentage points and increased the newcomer participation rate by 38% on average. Making community norms visible prevented unruly behavior within conversations; it also changed the group by influencing who chose to join.

Since individual preferences and societal structures contribute to gender inequality, efforts toward equality might need people to change their networks. This is especially true for journalists, whose source networks shape public knowledge of women's achievements. In the 1970s, Milton Rokeach found that confronting people with differences between their values and behavior might cause them to take steps toward equality. I tested this hypothesis with a field experiment and novel software on Twitter (FollowBias) that observed the percentage of women that journalists and bloggers followed, randomly assigned some to be confronted with that percentage, and suggested women to follow (n=139). While most who were confronted about the disparity expressed eagerness to follow more women, any effect was too small to observe with the sample size. I followed up with qualitative research on the forces against a person changing their network structure even when willing (Matias et al., 2017).

Can social movements change journalists' norms of reporting? Black Lives Matter campaigned to convince journalists and the public that unarmed black people killed by police were not isolated

incidents. In a paper under review, my collaborators and I test the hypothesis that journalists altered how they framed similar stories after the killing of Michael Brown in August 2014. Analyzing 11,114 news articles about 333 deaths from 2013-16, we found that an unarmed black person killed by the police received more news coverage after Michael Brown's death. The chance that a story mentioned at least one other victim also increased after his death. While coverage declined over time, the framing of these deaths as a systemic issue continued.

Future Directions on Behavior Change in Groups and Networks

In the next five years, I plan to conduct further field experiments that contribute pragmatic and theoretical knowledge on the roots and remedies of conflict, prejudice, and inequality.

While my research on preventing harassment has shown how social norms influence newcomer behavior, increased knowledge of norms could have different effects on people with a history of violating them. I am co-leading a US-wide study with Twitter that tests this and other hypotheses. As an industry-independent evaluation of a technology company policy, this study also pioneers the idea of behavioral consumer protection research (Benesch and Matias, 2018).

On reddit, communities with millions of subscribers expressed interest to conduct experiments after learning about my work. In January 2018, I convened the CivilServant Community Research Summit with representatives of 60 of the largest communities online. Working with researchers, communities developed new studies and replications. We have already begun several of these field experiments.

INTERDEPENDENT HUMAN AND MACHINE BEHAVIOR

Society now relies on automated systems to filter the information that informs human thoughts and actions. Governments and corporations also use algorithms to surveil behavior and enforce policies. Consequently, the work of maintaining democracies now involves managing algorithms as well as people. Because these algorithms both shape and respond to human behavior, attempts to theorize human behavior need to account for a world of adapting machines. In my research, I develop novel field experiments to observe and theorize the effects of human and algorithm behavior on each other.

When algorithms base their decisions on observations of humans, attempts to influence humans can influence those algorithms. In a field experiment (under review) testing an "AI nudge," I showed for the first time that an intervention can influence algorithm behavior by nudging human behavior In an online news discussion community of 14 million, I tested if encouraging readers to fact-check articles causes recommendation algorithms to behave differently. Interventions encouraged readers to (a) fact-check articles or (b) fact-check and vote to influence a recommendation algorithm. While both encouragements increased fact-checking behavior, only the fact-checking condition reduced an article's algorithmic ranking on average over time. Since AI nudges can influence algorithms, they have pragmatic and theoretical importance for understanding human and machine behavior.

Future Directions on Interdependent Human and Machine Behavior

In the next five years, I plan to continue work on theories and methods to understand and manage interdependent behavior of humans and machines.

Popularity algorithms routinely create conflict by amplifying contentious topics to large audiences. In a set of parallel field experiments underway with multi-million subscriber communities on the social

news platform reddit, I am testing the effect of community norm announcements that are targeted to conversations that have been amplified by algorithms. This research tests pragmatic ideas for reducing unruly behavior while also investigating algorithmic causes of conflict online.

AI systems that detect policy violations have enforced copyright law for over a decade. Qualitative research has found that knowledge of mass surveillance and automated law enforcement cause some people to withdraw from legitimate public discourse—a hypothesized "chilling effect." Competing theories attribute this effect to knowledge of surveillance or to fear of enforcement. U.S. federal courts have rejected both theories in civil liberties cases due to the weak state of evidence. To test these theories, colleagues and I are designing observational studies and field experiments with Twitter accounts that have received automated copyright enforcement.

Automated product testing systems routinely conduct thousands of concurrent behavioral studies, leading some to fear that mobile phones are creating automated addiction. In the Gray Phone Challenge, an n-of-one trial underway, colleagues and I have created software that supports people to discover if they experience these effects and test small changes that could reduce them. By combining personal treatment effects, we hope to discover if these findings generalize under what conditions.

CITIZEN BEHAVIORAL SCIENCE

Behavioral research can guide evidence-based uses of digital power. Unless this research is accountable to the public, it risks supporting new forms of authoritarianism. I do research and action toward democratic behavioral policy through innovations in citizen behavioral science, supporting the public to conduct research that holds power accountable and tests pragmatic ideas for change.

I first began thinking about citizen behavioral science while leading an audit of Twitter's responses to harassment in 2015 (Matias et al., 2015). The NGO Women, Action, and the Media (WAM!) asked me to analyze crowdsourced data on Twitter's policy responses to harassment cases. I realized that WAM!'s online safety work was similar to citizen science on the environment and food safety, a parallel I explored in articles for the Atlantic and the Guardian (Matias, 2015, 2016c).

A strike against reddit by volunteer moderators provided my first opportunity to develop participatory hypothesis testing as a method. While political scientists have theorized factors predicting participation in social movements, theories tend to originate with researchers rather than movements. In interviews and large-scale public discussions, I crowdsourced predictors and explanations for strike participation and incorporated them into a model that tests participant theories across 52,735 communities. Together we discovered factors including community grievances, resources, social isolation, and elite networks that predicted a group's participation in the strike (Matias, 2016b).

During my ethnographic research on volunteer online governance (Matias, 2016a), communities often requested knowledge about the outcomes of community interventions and technology company policies. At the time, I was also researching early figures of behavioral policy including Donald Campbell and Karl Popper, who feared that social experiments would advance authoritarian power.

Drawing from my ethnographic fieldwork and historical research, I developed CivilServant, software that supports communities to lead their own behavioral experiments independently from tech companies (Matias and Mou, 2018). CivilServant is now a nonprofit with three staff. Incubated by the NGO Global Voices, CivilServant collaborates with the public in behavioral science for a flourishing internet. We expect our network of researchers and communities to complete ten new studies in 2018.

Future Directions on Citizen Behavioral Science

In the next five years, I plan to scale citizen behavioral science by broadening who participates, developing methodological advancements, and innovating on the ethics of large-scale online research.

This year, CivilServant is extending to Twitter, Wikipedia, and mobile phones, supported by grants from the Ethics & Governance of AI Fund and the Templeton World Charity Foundation. We have also received funding from the Knight Foundation, the MacArthur Foundation, the Mozilla Foundation, and the Tow Center for Digital Journalism at Columbia University.

Responsibly broadening the public's capacity to conduct behavioral science requires innovations in research ethics, an area where I am conducting empirical research. For example, in cases where participant knowledge of research might weaken validity, novel polling methods may achieve workable forms of consent and participant agency. In a recent paper, my coauthor and I present an automated system that (a) solicits ethics feedback with a representative sample of a study population and (b) debriefs participants, giving them options to manage their data privacy (Zong and Matias, 2018).

SUMMARY

Digital communications continue to restructure how we relate to others in groups, networks, and as societies. My research has shown how norms influence behavior online and how people and movements organize to change norms and structures. As algorithms grow in power and pervasiveness, I have shown that we can change how those algorithms behave by influencing human behavior. Each of these studies has developed pragmatic knowledge affecting millions of people. Throughout my work, I have applied historical and ethical lenses to re-imagine the role of behavioral science in democracy and to create novel methods for holding digital power accountable to the public. I am excited to continue advancing a public-interest research program that interlinks theory, methods, and practice.

References

S. Benesch and J. N. Matias. Launching today: new collaborative study to diminish abuse on Twitter, Apr. 2018. URL https://medium.com/@susanbenesch/launching-today-new-collaborative-study-to-diminish-abuse-on-twitter-2b91837668cc.

K. Lewin. *Resolving social conflicts; selected papers on group dynamics*. Harper & Row, 1948.

J. Matias, A. Johnson, W. E. Boesel, B. Keegan, J. Friedman, and C. DeTar. Reporting, reviewing, and responding to harassment on Twitter. 2015. URL https://arxiv.org/abs/1505.03359.

J. N. Matias. The Tragedy of the Digital Commons: Advocates for fairer, safer online spaces are turning to the conservation movement for inspiration. *The Atlantic*, June 2015. ISSN 1072-7825. URL http://www.theatlantic.com/technology/archive/2015/06/the-tragedy-of-the-digital-commons/395129/.

J. N. Matias. The Civic Labor of Online Moderators. In *Internet Politics and Policy conference, Oxford, United Kingdom*, 2016a.

J. N. Matias. 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*, pages 1138–1151. ACM, 2016b.

J. N. Matias. A toxic web: what the Victorians can teach us about online abuse. *The Guardian*, Apr. 2016c. URL https://www.theguardian.com/technology/2016/apr/18/a-toxic-web-what-the-victorians-can-teach-us-about-online-abuse.

J. N. Matias and M. Mou. CivilServant: Community-Led Experiments in Platform Governance. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, page 9. ACM, 2018.

J. N. Matias, S. Szalavitz, and E. Zuckerman. FollowBias: Supporting Behavior Change toward Gender Equality by Networked Gatekeepers on Social Media. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pages 1082–1095. ACM, 2017.

J. Zong and J. N. Matias. Automated Debriefing: Interface for Large-Scale Research Ethics. In *2018 ACM Conference on Computer Supported Cooperative Work and Social Computing*, June 2018.