Preventing harassment and increasing group participation through social norms in 2,190 online science discussions

J. Nathan Matias[a,b,1]

[a]Department of Psychology, Princeton University, Princeton, NJ 08540; and [b]Center for Information Technology Policy, Princeton University, Princeton, NJ 08540

Theories of human behavior suggest that people's decisions to join a group and their subsequent behavior are influenced by perceptions of what is socially normative. In online discussions, where unruly, harassing behavior is common, displaying community rules could reduce concerns about harassment that prevent people from joining while also influencing the behavior of those who do participate. An experiment tested these theories by randomizing announcements of community rules to large-scale online conversations in a science-discussion community with 13 million subscribers. Compared with discussions with no mention of community expectations, displaying the rules increased newcomer rule compliance by >8 percentage points and increased the participation rate of newcomers in discussions by 70% on average. Making community norms visible prevented unruly and harassing conversations by influencing how people behaved within the conversation and also by influencing who chose to join.

online harassment | group participation | social norms | field experiment | science communications

How do a community's rules about behavior influence people's decisions to participate in an online conversation and how they subsequently behave? Questions about newcomer participation are part of a broader scientific discussion about how groups develop, grow, and cultivate shared patterns of behavior (1). These questions are pragmatically important in online communications, where rapidly forming conversations and popularity algorithms can attract tens of thousands of people into settings where they experience unruly and harassing behavior (2, 3).

Research on social norms has discovered how human behavior is guided by our subjective perceptions about what is common or acceptable to others in a situation. When forming these perceptions, we choose which sources of information about norms to attend to, especially norms from "reference groups" that are important to us (4). Influential signals of observed behavior and described norms can come from individual referents (5, 6) or from groups (7). Institutions also shape norm perceptions (8), especially when institutions indicate how those expectations will be enforced (9, 10).

Because newcomers to a group have limited ties to the group, theories of social norms predict that information about norms may not influence their behavior. Numerous studies have found that information about norms is influential when a person cares about a group or they feel like they belong (11). Yet scientists have long observed that people are more likely to follow perceived norms when their behavior is less private (12, 13). In online conversations, normative messages might still influence newcomer behavior when people know they are monitored and norms are continuously enforced (14).

Even if signals of group norms are less influential among individual newcomers, they might still influence groups through a process of selection. Scientists have described the decision to join a group as a process of reconnaissance and evaluation during which a person discovers information about a group and makes judgments about whether to join (1, 15). Might information about social norms influence group behavior by influencing a newcomer's decision to join the group?

Researchers have tended to study decisions about group participation in slow, small processes where selection is costly, such as joining a club or accepting a job offer. Joining often involves a two-sided process of approval from the group and the newcomer (15, 16). Similarly, research on the effect of social norms has investigated settings where the cost of norm compliance is lower than the cost of leaving. Perhaps signs in parking garages can reduce littering because few people would choose a different garage to avoid using a trash bin as instructed (7, 9).

In text-based online conversations, the choice to join or leave a conversation is faster and lower-cost than joining a club or leaving a parking garage. On social platforms like Facebook and Reddit, people make frequent choices about what conversations to join in unfamiliar settings. These platforms host hundreds of thousands of parallel communities and continuously promote conversations from them to tens of millions of readers—people who choose again and again where to allocate their attention and voice (17). On these platforms, joining a group involves minimal effort and avoiding one includes no cost at all, since the group never knows you were there.

Significance

Online harassment remains a common experience despite decades of work to identify unruly behavior and enforce rules against it. Consequently, many people avoid participating in online conversations for fear of harassment. Using a large-scale field experiment in a community with 13 million subscribers, I show that it is possible to prevent unruly behavior and also increase newcomer participation in public discussions of science. Announcements of community rules in discussions increased the chance of rule compliance by >8 percentage points and increased newcomer participation by 70% on average. This study demonstrates the influence of community rules on who chooses to join a group and how they behave.

PSYCHOLOGICAL AND COGNITIVE SCIENCES

Fig. 1. Treatment messages posted to discussions of peer-reviewed research.

In these online groups, information about social norms may influence a person's decision to join a conversation by informing how they expect to be treated by others. Many people currently avoid speaking publicly online for fear of harassment. In the United States, 47% of internet users report experiencing some kind of online harassment, with 20% of victims choosing to shut down an online account to protect themselves (18). Groups that systematically receive the most severe forms of harassment report in surveys that they would be more likely to participate online in the presence of national laws on harassment and cyberbullying (19). Furthermore, laboratory experiments comparing discussion-software designs have found that readers express a greater intent to comment in conversation environments that include continuous monitoring and enforcement of moderation policies (20).

Using a large-scale field experiment, I tested the hypothesis that normative information about a community's rules influences newcomers' choices to participate and how they behave. I designed this experiment with The New Reddit Journal of Science (r/science), a 13-million-subscriber community on the Reddit platform. The community hosts discussions about peer-reviewed journal articles and live question-and-answer (Q&A) sessions with prominent scientists. Many of these discussions attract conflict and harassment. For example, when Professor Stephen Hawking answered questions in 2015, commenters mocked his medical condition and personal life with abusive and obscene insults. In July 2016, when discussing new research about obesity among women, moderators removed 1,497 of 2,186 comments. Many of those comments derided women, mocked the researchers, and criticized people with obesity through jokes, memes, personal anecdotes, and unsubstantiated medical advice. Conversations about politically charged topics—including race, gender, vaccines, and climate science—also attract substantial harassment. That month, moderators removed >35,000 comments or discussions and banned 460 people from future participation for violating community rules.

Volunteer-led online communities like r/science are fruitful settings to investigate questions about online behavior because these communities carry out much of the policy-making and enforcement on the internet. When someone threatens, disparages, or otherwise harasses another person online, volunteer community moderators offer one of three kinds of authority governing that behavior. Government regulations about hate speech or threats of violence sometimes lead to court cases and content removals (21, 22). The operators of online platforms such as Facebook, Google, and Reddit also develop their own policies that are enforced by thousands of paid staff (23–29). Yet platforms struggle to develop scalable policies that can be consistently applied across jurisdiction, language, context, and culture for billions of people (30). To meet these context-specific needs, hundreds of thousands of volunteer community moderators provide the most local form of governance for communities that sometimes grow to tens of millions of subscribers (31). These volunteer teams create rules (32), monitor activity (33), carry out policy interventions, and report serious cases to platforms and law enforcement (34).

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 of the comments they removed on average and 52.3% of all newcomer comments. First-time commenters were also more likely to violate community policies than more experienced commenters. These newcomers may not yet be aware of community policies against abusive language, insulting jokes, or personal medical anecdotes. Since newcomers are important for the community's science-communication goals and are also a major source of unruly behavior, the community has a pragmatic interest in efforts to influence who participates and how they behave.

Because these community moderators create and administer policy for millions of people in data-rich online environments (34), moderators have unique opportunities to conduct field experiments on social norms. In the social sciences, researchers have used field experiments to validate existing theories or identify new phenomena and hypotheses for further investigation (35, 36). In this study, community moderators identified the problem of newcomer rule compliance and advised on intervention design, study procedures, and outcomes.

Moderators decided to test a "sticky comment" that displayed community rules at the top of a discussion. We designed the intervention to include information about norms that other research has found effective at influencing behavior (4). The message welcomes participants, names the unacceptable behavior, describes the enforcement consequences, reports that many people agree with the norm, and indicates the community's capacity to monitor and enforce its policies (Fig. 1).

The experiment was conducted by software that observed new discussions as they were posted and determined whether they focused on an article or live Q&A. The software then randomly assigned a discussion to receive an announcement displaying rules of participation or no message at all. In discussions that received this intervention, every person who read and commented in the discussion was shown the announcement. Throughout the study, >1,000 moderators, who were blinded from the experiment conditions, continued their practice of removing comments that violated community rules. Over 30 d, the software observed discussions, observed comments, counted the number of first-time commenters, and recorded which of the comments were removed by moderators.

Results

I evaluated the effect of rule postings in r/science from August 25, 2016, through September 23, 2016. The experiment included 2,190 discussions of academic publications. The 18,264 newcomer comments were 29% of all comments in this period and were made in 804 of the discussions in the experiment.

In this science-discussion community, posting the rules caused newcomer comments in discussions to comply with community rules at higher rates ($P = 0.038$, Table 1 and Fig. 2). Without posting the rules, a first-time commenter in a discussion about a peer-reviewed article has a 52.5% chance of complying with

Table 1. Random intercepts logistic regression estimating the average treatment effect of posting rules to discussions on the chance of newcomer comment rule compliance

	Rule compliance
TREAT	0.34* (0.16)
(Intercept)	0.10 (0.12)
Log likelihood	−10,772.37
Number observed	18,264
Number of groups: link$_i$d	804
Variable: link$_i$d(*Intercept*)	2.66

*$P < 0.05$.

A

Effect on Log−Odds of Newcomer Rule Compliance

Effect of posting rules on the log−odds of a newcomer comment to be removed
by moderators in r/science community from 08/25/2016 − 09/23/2016.
n = 62759 newcomers, 1845 posts, in a random intercepts logistic regression (p=0.0446).

B

Effect on Incidence Rate of Newcomer Comments

Effect of posting rules on incidence rate of number of newcomer comments
by moderators in r/science community from 08/25/2016 − 09/23/2016.
n = 62759 newcomers, 2190 posts, in a negative binomial model (p < 0.001).

C

Effect on Incidence Rate of Newcomer Comment Removals

Effect of posting rules on incidence rate of number of newcomer comments removed
by moderators in r/science community from 08/25/2016 − 09/23/2016.
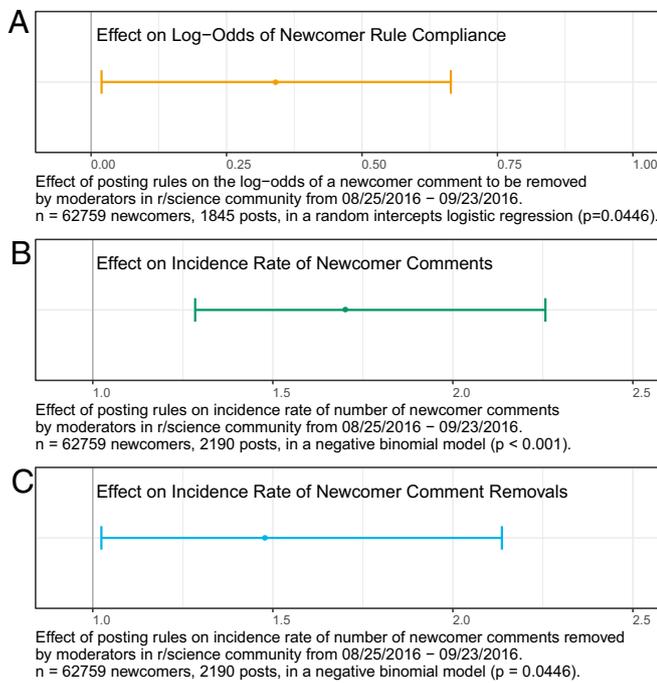n = 62759 newcomers, 2190 posts, in a negative binomial model (p = 0.0446).

Fig. 2. (*A* and *B*) Posting messages announcing rules increased the chance of individual rule compliance by newcomers (*A*) and increased the rate of newcomer participation (*B*). (*C*) With this increase in participation, the rate of comment removals also increased.

community norms. Posting the rules causes an 8.4-percentage-point increase in the chance that a newcomer's comment will be allowed to remain by moderators on average in r/science.

Posting the rules also affects who chooses to participate in a conversation, increasing the participation rate of first-time commenters in a discussion by 70% on average ($P < 0.0001$; Table 2 and Fig. 2*B*). Because newcomers participate at higher rates, the intervention may also create more work for moderators, increasing the rate of newcomer comments removed on average ($P < 0.0001$; Table 2 and Fig. 2*C*).

Discussion

Posting community rules to discussions influenced who joined the discussion and how they behaved. Despite theoretical expectations that newcomers might not be influenced by normative information about communities to which they do not belong, the intervention increased norm compliance among first-time participants. This outcome may be due to the announcement's specific descriptions of community expectations, enforcement details, information on the number of participants who ascribe to the norm, and suggestion of continuous enforcement.

In an online setting where joining conversations is a fast decision and harassment is a real risk, the intervention also caused newcomers to participate in discussions at higher rates. Social norms don't just affect the behavior of people who already participate in a social context. By providing information about how other people will behave, they also have a strong effect on a newcomer's decision to participate in a community for the first time.

While this field experiment demonstrates the pragmatic effects of social-norms interventions on behavior, it does not distinguish discussion-level effects from individual ones. Because the intervention was applied to discussions and because the software could not observe people who viewed the intervention and decided not to contribute, it is not possible to disentangle the decision to participate from the decision about how to participate. Since the intervention likely attracted some first-time

commenters who were already more likely to comply with the rules, any effects on behavior should be interpreted as closely associated with the effects on the decision to participate. Furthermore, this study cannot offer evidence on whether this intervention reduced harassing behaviors overall or displaced it to other communities. This finding is also likely sensitive to contextual factors about a given online community, the pool of potential participants, and the design of the software platform.

People who decline to discuss science in public may have good reasons to hold back, given the rate of online harassment. Even as technology companies attempt to respond to unruly behavior after it occurs, volunteer community moderators can also manage and prevent problems for thousands of people in their own communities. Pragmatically, in the r/science community, displaying the rules could prevent >2,000 first-time commenters per month from unruly behavior and increase first-time commenters on scientific topics by >40,000 people per month on average. Experiments by and with real-world communities can evaluate such interventions while also helping social scientists understand how communities form and behave online.

Materials and Methods

I conducted this experiment with the CivilServant software, which supports community-based experiments on the Reddit platform (37). The software was granted moderator privileges by moderators of r/science independently of the Reddit company, giving it access to make observations about new discussions, automatically post randomly assigned announcements to those discussions, observe commenter participation, and incorporate administrative records on which comments were removed by moderators for violating community policies.

Data Collection. For each discussion, the software observed information about the discussion, comments in the discussion, and the user accounts that made those comments. All discussions in the r/science Subreddit during the experiment period were automatically observed by the experiment software and assigned a condition after they were started by community members, except in cases of spam auto-removed by the Reddit platform before the experiment software could make an observation. Comments were included in the study if they were made in an observed discussion. Among participants, newcomers were accounts that had not previously contributed in the community in the previous 6 mo before the study began. To determine newcomer status, I used a list of all accounts that had previously contributed public comments to the community.

The main unit of observation for this study was a comment by a newcomer in a discussion within the experiment sample. In the life cycle of a comment, it may be removed by moderators or an automated moderation bot. This removal action may be reverted by moderators, who can make the comment visible again, sometimes after deliberation about the case in question. This research considered the final visibility of the comment after the experiment concluded. Since moderators were blinded from the treatment and consequently could not review direct replies to the treatment announcement, the software automatically removed all replies to the treatment announcement and omitted them from the study. Discussion-level outcomes for the study included the total number of first comments by newcomers in a discussion, as well as the total number of newcomer comments removed by moderators in a discussion.

Table 2. Estimates of average treatment effects among newcomers in article discussions

Estimator	Estimate	Standard error	P	P (adjusted)
Number of comments				
NegBin	0.532***	0.1438	0.0002	0.0009
RI	4.336*	NA	0.0364	0.0446
Number removed				
NegBin	0.391*	0.1876	0.0370	0.0446
RI	1.687	NA	0.1134	0.1228

Discussions = 2,190; newcomer comments = 18,264. RI estimates are one-tailed randomization inference tests. NA, not applicable.
*$P < 0.05$; ***$P < 0.001$ (adjusted for 13 comparisons; Benjamini–Hochberg).

Treatment Assignment. During the experiment, the software observed every new discussion as it was posted and determined whether or not it discussed a peer-reviewed article or was a Q&A with a scientist (like the discussion with Stephen Hawking). Treatments were randomized within each type of discussion and also block-randomized over time (blocks of 6 for Q&A discussions and blocks of 10 for articles).

Informed consent was waived for this study, which involved months of community consultation and approval. I debriefed the community about the study conclusions in a community-wide announcement visible to anyone visiting r/science. This research was approved by the MIT Institutional Review Board (protocol 1604554325).

As specified in the preanalysis plan (https://osf.io/knb48/files/osfstorage/57bef819594d9001fcd0e193/), I ended the study according to the stop rule after analyzing data from 2,200 assignments, since the estimated effect on the incidence rate of newcomer comments removed was a >20% increase (Table 2). After removing 24 discussions from randomization blocks that were spoiled by software errors, the final results included a total of 24 Q&A discussions and 2,190 academic-article discussions.

This paper focuses on the experiment within the 2,190 discussions of scientific articles, distinct from the Q&A discussions with scientists. Among these article discussions, 804 included at least one newcomer comment.

Analysis. I conducted a series of logistic regression and negative binomial models to test the effect of normative messages on the rule-compliance and participation rates of newcomers in discussions of scientific articles. The decision rule for these estimates is $\alpha = 0.05$, with P values adjusted for multiple comparisons with the Benjamini–Hochberg method (38).

I estimated the average treatment effect on the chance of a newcomer comment's rule compliance using a logistic regression model with random intercepts for the discussion that was treated (μ_j). By using a multilevel model, I was able to estimate the effect on comments when the treatment was allocated to discussions.

$$logit(\pi_i) = \beta_0 + \beta_1 TREAT_{ij} + \mu_j + \epsilon_{ij}.$$

I estimated the average treatment effect on the incidence rates of newcomer comments and removed comments per discussion with negative binomial models. These models offer estimates on the rate of occurrences in settings like comment discussions where higher counts of incidents attract even more incidents (39).

$$log(\mu_i) = \beta_0 + \beta_1 TREAT_i + \epsilon_i.$$

I also included two randomization inference tests on a sharp null of no effect on the count of newcomer comments and removed newcomer comments, using the randomization procedures from the study design (40).

On average, in discussions of scientific publications, I found that posting rules to a discussion increased the chance of a newcomer comment to comply with the rules by 8.4 percentage points ($P = 0.014$) (Table 1). Posting the rules also increased the rate of newcomer comments by 70% ($P < 0.0001$). This increase in newcomer participation was consistent with the results of the randomization inference test ($P = 0.014$). While an increase in newcomer participation could also create a greater burden for moderators on average, the results of that hypothesis are inconclusive. In a negative binomial model, I found an increase of 48% in newcomer comments removed ($P = 0.0487$), but I failed to reject the sharp null from randomization inference ($P = 0.121$). Full results from every estimator are reported in Table 2.

Since this paper reports a subgroup analysis of a larger experiment conducted with r/science, all P values were adjusted as part of a family of 13 hypothesis tests, including the preregistered analyses, which were consistent with these findings. A full report of these analyses is included in *SI Appendix*.

Statement on Data. The experimental design was preregistered at https://osf.io/knb48/files/osfstorage/57bef819594d9001fcd0e193/. Code and data for this paper are available on Github (41).

1. Levine JM, Moreland RL (1994) Group socialization: Theory and research. *Eur Rev Soc Psychol* 5:305–336.
2. Massanari A (2015) #Gamergate and the fappening: How Reddit's algorithm, governance, and culture support toxic technocultures. *New Media Soc* 19:329–346.
3. Margetts H, John P, Hale S, Yasseri T (2015) *Political Turbulence: How Social Media Shape Collective Action* (Princeton Univ Press, Princeton).
4. Tankard ME, Paluck EL (2016) Norm perception as a vehicle for social change. *Soc Issues Policy Rev* 10:181–211.
5. Cialdini RB, Kallgren CA, Reno RR (1991) A focus theory of normative conduct: A theoretical refinement and reevaluation of the role of norms in human behavior. *Adv Exp Soc Psychol* 24:1–243.
6. Paluck EL, Shepherd H, Aronow PM (2016) Changing climates of conflict: A social network experiment in 56 schools. *Proc Natl Acad Sci USA* 113:566–571.
7. Goldstein NJ, Cialdini RB, Griskevicius V (2008) A room with a viewpoint: Using social norms to motivate environmental conservation in hotels. *J Consum Res* 35:472–482.
8. Cialdini RB, Goldstein NJ (2004) Social influence: Compliance and conformity. *Annu Rev Psychol* 55:591–621.
9. Reiter SM, Samuel W (1980) Littering as a function of prior litter and the presence or absence of prohibitive signs. *J Appl Soc Psychol* 10:45–55.
10. De Kort YA, McCalley LT, Midden CJ (2008) Persuasive trash cans: Activation of littering norms by design. *Environ Behav* 40:870–891.
11. Perkins HW (2002) Social norms and the prevention of alcohol misuse in collegiate contexts. *J Stud Alcohol*, 164–172.
12. Deutsch M, Gerard HB (1955) A study of normative and informational social influences upon individual judgment. *J Abnorm Soc Psychol* 51:629–636.
13. Hogg MA (2010) Influence and leadership. *Handbook of Social Psychology*, eds Fiske ST, Gilbert DT, Lindzey G (John Wiley & Sons, Inc., New York), pp 1166–1207.
14. Zittrain J (2007) Perfect enforcement on tomorrow's internet. *Regulating Technologies: Legal Futures, Regulatory Frames and Technological Fixes*, eds Brownsword R, Yeung K (Hart, Oxford), pp 125–156.
15. Pavelchak MA, Moreland RL, Levine JM (1986) Effects of prior group memberships on subsequent reconnaissance activities. *J Personal Soc Psychol* 50:56–66.
16. Chatman JA (1989) Matching people and organizations: Selection and socialization in public accounting firms. *Academy of Management Proceedings* (Academy of Management, Briarcliff Manor, NY), Vol 1989, pp 199–203.
17. Massanari AL (2015) *Participatory Culture, Community, and Play: Learning from Reddit, Digital Formations* (Peter Lang, New York).
18. Lenhart A, Ybarra M, Zickuhr K, Price-Feeney M (2016) Online harassment, digital abuse, and cyberstalking in America (Data & Society Research Institute, New York), Technical Report.
19. Penney JW (May 26, 2017) Internet surveillance, regulation, and chilling effects online: A comparative case study. *Internet Policy Rev*, 10.14763/2017.2.692.
20. Wise K, Hamman B, Thorson K (2006) Moderation, response rate, and message interactivity: Features of online communities and their effects on intent to participate. *J Comput Mediated Commun* 12:24–41.
21. Citron DK (2009) Law's expressive value in combating cyber gender harassment *Mich L Rev* 108:373–415.
22. Marwick AE, Miller RW (2014) *Online harassment, defamation, and hateful speech: A primer of the legal landscape* (Social Science Research Network, Rochester, NY), SSRN Scholarly Paper ID 2447904.
23. Gillespie T (2018) *Custodians of the Internet: Platforms, Content Moderation, and the Hidden Decisions that Shape Social Media* (Yale Univ Press, New Haven).
24. Citron DK, Norton HL (2011) Intermediaries and hate speech: Fostering digital citizenship for our information age. *Boston Univ Law Rev* 91:1435.
25. Chen A (2014) The laborers who keep dick pics and beheadings out of your Facebook feed. Wired. Available at https://www.wired.com/2014/10/content-moderation/. Accessed February 7, 2017.
26. Crawford K, Gillespie TL (2014) What is a flag for? Social media reporting tools and the vocabulary of complaint. *New Media Soc* 18:410–428.
27. Matias JN, et al. (2015) Reporting, reviewing, and responding to harassment on Twitter. arXiv:1505.03359. Preprint, deposited May 13, 2015.
28. Roberts ST (2016) Commercial content moderation: Digital laborers' dirty work. *The Intersectional Internet: Race, Sex, Class, and Culture Online*, eds Noble SU, Tynes BM (Peter Lang, New York), pp 147–161.
29. Buni C, Chemaly S (2016) *The secret rules of the Internet*. The Verge. Available at https://www.theverge.com/2016/4/13/11387934/internet-moderator-history-youtube-facebook-reddit-censorship-free-speech. Accessed April 26, 2017.
30. MacKinnon R (2013) *Consent of the Networked: The Worldwide Struggle for Internet Freedom* (Basic Books, New York), Reprint ed.
31. Seering J, Wang T, Yoon J, Kaufman G (January 11, 2019) Moderator engagement and community development in the age of algorithms. *New Media Soc*, 10.1177/1461444818822131.
32. 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*, ed Weisband S (Routledge, New York), pp 171–194.
33. Geiger RS, Ribes D (2010) The work of sustaining order in Wikipedia: The banning of a vandal. *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work* (Association for Computing Machinery, New York), pp 117–126.
34. Matias JN (2019) The civic labor of volunteer moderators online. *Soc Media Soc*, 10.1177/2056305119836778.
35. Cialdini RB (1980) Full-cycle social psychology. *Appl Soc Psychol Annu* 1:21–47.

36. Paluck EL, Cialdini RB (2014) Field research methods. *Handbook of Research Methods in Social and Personality Psychology*, eds Reis HT, Judd CM (Cambridge Univ Press, New York), pp 81–97.
37. Matias JN, Mou M (2018) CivilServant: Community-led experiments in platform governance. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Association for Computing Machinery, New York), p 9.
38. Benjamini Y, Hochberg Y (1995) Controlling the false discovery rate: A practical and powerful approach to multiple testing. *J R Stat Soc Ser B Stat Methodol* 57:289–300.
39. Long JS, Freese J (2014) *Regression Models for Categorical Dependent Variables Using Stata* (Stata, College Station, TX), 3rd Ed.
40. Coppock A (2019) Randomization inference with ri2 (Yale University, New Haven), Technical Report.
41. Matias JN (2019) Data from "Estimating the effect of public postings of norms in subreddits: Data and code." Github. Available at https://github.com/mitmedialab/CivilServant-Analysis/blob/master/papers/r_science_2016/. Deposited February 1, 2019.