Citation:
Brady, W.J., McLoughlin, K., Doan, T.N., & Crockett, M.J. (2021). How social learning amplifies moral outrage expression in online social networks. *Science Advances*, 7(33), eabe5641.

Link to print version: https://advances.sciencemag.org/content/7/33/eabe5641

SUPPLEMENTARY ONLINE MATERIALS (SOM) ARE AVAILABLE HERE

FRONT MATTER

Title

How social learning amplifies moral outrage expression in online social networks

Short title: Social learning of outrage in online networks

Authors

William J. Brady,[1]* Killian McLoughlin[1], Tuan N. Doan[2], & Molly J. Crockett[1]*

Affiliations

[1]Yale University, Department of Psychology
[2]Yale University, Department of Statistics and Data Science
* Correspondence to William J. Brady & M.J. Crockett
Email: william.brady@yale.edu, mj.crockett@yale.edu

Abstract

Moral outrage shapes fundamental aspects of social life and is now widespread in online social networks. Here, we show how social learning processes amplify online moral outrage expressions over time. In two preregistered observational studies on Twitter (7331 users and 12.7 million total tweets) and two preregistered behavioral experiments (N = 240), we find that positive social feedback for outrage expressions increases the likelihood of future outrage expressions, consistent with principles of reinforcement learning. In addition, users conform their outrage expressions to the expressive norms of their social networks, suggesting norm learning also guides online outrage expressions. Norm learning overshadows reinforcement learning when normative information is readily observable: in ideologically extreme networks, where outrage expression is more common, users are less sensitive to social feedback when deciding whether to express outrage. Our findings highlight how platform design interacts with human learning mechanisms to affect moral discourse in digital public spaces.

MAIN TEXT

Introduction

Moral outrage is a powerful emotion with important consequences for society (*1–3*): it motivates punishment of moral transgressions (*4*), promotes social cooperation (*5*) and catalyzes collective action for social change (*6*). At the same time, moral outrage has recently been blamed for a host of social ills, including the rise of political polarization (*7, 8*), the chilling of public speech (*9*), the spreading of disinformation (*10*), and the erosion of democracy (*11*). Some have speculated that social media can exacerbate these problems by amplifying moral outrage (*11*). However, evidence to support such claims remains scarce. Our current understanding of moral outrage is largely based on studies examining its function in small group settings (*2, 12*), which impose very different constraints on behavior than online environments (*13, 14*). There is therefore a pressing need to understand the nature of moral outrage as it unfolds in online social networks.

Foundational research shows that people experience moral outrage when they perceive a moral norm has been violated (*2, 15–17*), and express outrage when they believe it will prevent future violations (*5, 18*) or promote social justice more broadly (*6*). At the same time, however, outrage expressions may be sensitive to factors that have less to do with individual moral convictions, particularly in the context of social media. More specifically, we suggest that online outrage expressions are shaped by two distinct forms of learning. First, people may change their outrage expressions over time through *reinforcement learning*, altering expressive behaviors in response to positive or negative social feedback (*13, 19, 20*). Second, people may adjust their outrage expressions through *norm learning*, matching their expressions to what they infer is normative among their peers through observation (*21–25*). Social media platforms have specific design features that can impact both forms of learning: they deliver highly salient, quantifiable social feedback (in the form of 'likes' and 'shares'), a central component of reinforcement learning; and they enable users to self-organize into homophilic social networks with their own local norms of expression displayed in newsfeeds (*26, 27*), which should guide norm learning.

Supporting these hypotheses, recent work demonstrates that social media users post more frequently after receiving positive social feedback (*28*), consistent with a reinforcement learning account. These observations lead to a straightforward prediction that social media users' current moral outrage expressions should be positively predicted by the social feedback ('likes' and 'shares') they received when they expressed moral outrage in the past. Furthermore, because moral and emotional expressions like outrage receive especially high levels of social feedback (*29–31*), moral outrage expressions may be especially likely to increase over time via social reinforcement learning. Finding evidence for this would contradict the idea that social media platforms provide neutral channels for social expressions and do not alter those expressions.

However, reinforcement learning alone is unlikely to fully explain the dynamics of online moral outrage expression. Social media users interact with others in large social networks, each with its own norms of expression (*27*). Every time a user logs onto a platform, their newsfeed immediately provides a snapshot of the communication norms currently present in their network (*26*). This information is likely to guide norm learning, where users adjust their behavior by following what others do, rather than responding to reinforcement (*21, 22, 32–36*). Crucially, reinforcement learning and norm learning processes might interact with one another: when individuals can directly observe which actions are most valuable, they rely less on reinforcement learning (*22, 37*). Thus, moral outrage expressions might be guided more by norm learning than reinforcement learning when normative information is readily observable in a network.

We tested our hypotheses across two pre-registered observational studies of Twitter users, and two pre-registered behavioral experiments in a simulated Twitter environment. Collectively, this work demonstrates that social media users' moral outrage expressions are sensitive to both

direct social feedback and network-level norms of expression. These findings illustrate how the interaction of human psychology and digital platform design can impact moral behavior in the digital age (*26*, *35*, *38*, *39*).

Results

Studies 1 and 2

Measuring moral outrage

To test our hypotheses, we developed a method for measuring moral outrage expressions at scale in social media text, focusing on Twitter as our data source. This platform is appropriate for testing our hypotheses due to the occurrence of several high-profile, rapid swells of outrage on this platform (*40*) and the fact that many important public figures use it to communicate with their audiences, frequently expressing and provoking outrage both online and offline. We used supervised machine learning to develop a *Digital Outrage Classifier* (DOC; **Materials and Methods**) that can classify tweets as containing moral outrage or not. To train DOC, we collected a set of 26,000 tweets from a variety of episodes that sparked widespread public outrage (see **Materials and Methods** and **Table 1**), and used theoretical insights from social psychology to annotate those tweets according to whether they expressed moral outrage. The key definition of moral outrage included the following three components (*1*, *2*, *41*): a person can be viewed as expressing moral outrage if (a) they have feelings in response to a perceived violation of their personal morals, (b) their feelings are comprised of emotions such as anger, disgust and contempt, and (c) the feelings are associated with specific reactions including blaming people/events/things, holding them responsible, or wanting to punish them. The full instructions including examples given to participants and distinctions between moral outrage and other related concepts (e.g., "pure trolling") can be viewed in SOM, Section 1.2.

To enhance generalizability of our classifier, our annotated dataset contained episodes that spanned diverse topics, ideologies and timepoints. **Table 2** provides examples of classifications made by DOC. Extensive evaluation demonstrated that DOC classified moral outrage in tweets with reliability comparable to expert human annotators (see **Materials and Methods**). DOC is freely available for academic researchers via a Python package at the following link: https://github.com/CrockettLab/outrage_classifier.

Topic	Description	Tweet Date Range	Political Ideology of Users	Tweets containing outrage	*N*
Kavanaugh	During the confirmation process for the Supreme Court, nominee Brett Kavanaugh was accused of sexually assaulting Dr. Christine Blasey Ford. Both parties testified to the Senate Judiciary Committee, and Kavanaugh was ultimately confirmed.	Sep 15 – Oct 18, 2018	Mixed	52.00%	16,000
Covington	White high school students wearing "Make America Great Again" hats were filmed appearing to harass a Native American man in Washington, D.C.	Jan 22 - Feb 1, 2019	Mixed	26.36%	2,500

	After the video went viral, subsequent footage suggested that the interaction was more complicated. Several media outlets issued retractions.				
United	A United Airlines passenger was forcibly removed from an overbooked plane. Footage of the event showed the passenger being injured. The video went viral and elicited backlash against the airline.	Apr 10 - 14, 2017	Mixed	20.08%	2,500
Smollett	In January 2019, actor Jussie Smollett claimed to have been the victim of a violent hate crime perpetrated by supporters of President Trump. Investigating officers later alleged in February that Smollett had staged the attack.	Feb 22 - 26, 2019	Conservative	23.00%	2,500
Transgender Ban	The Trump administration's ban on transgender individuals serving in the military was upheld by the US Supreme Court, reversing the 2016 decision by President Obama to open the military to transgender service members.	Jan 22 - 25, 2019	Liberal	52.60%	2,500

Table 1. Characteristics of all training datasets. DOC was first trained on 16,000 tweets collected during the Brett Kavanaugh confirmation hearings. We then tested generalizability and re-trained on the combination of Kavanaugh and all other topics (26,000 total tweets).

Our measurement of moral outrage is based on a theoretical assumption that it is a specific subcategory of the broader category of negative sentiment, which in addition to moral outrage includes other negative emotion expressions such as fear and sadness (*2*, *42*). In other words, we expected that expressions of negative sentiment are necessary but not sufficient for positive classifications by DOC. We examined this expectation by testing DOC's discriminant validity against a negative sentiment classifier (NSC) trained on the widely-used Sentiment140 dataset (*43*). We predicted that DOC's and the NSC's classifications would be correlated but would also have many cases of non-overlap. To test this prediction, we analyzed our 26,000-tweet dataset used to train DOC (described in Table 1) to compare moral outrage classifications by DOC and negative sentiment classifications by the NSC. As expected, we found a weak correlation between the two classifiers' outputs using Kendall's rank correlation test, $\tau = .11$, $p < .001$. Thus, we demonstrate discriminant validity: DOC's classifications and the NSC's classifications are correlated, but not identical. See SI Appendix, Section 1.7 for more details.

Reinforcement learning hypothesis

Our first hypothesis was that positive social feedback for previous outrage expressions should predict subsequent outrage expressions. To test this, we used Twitter's standard and premium APIs to collect the full tweet histories of 3669 "politically engaged" users who tweeted at least once about the Brett Kavanaugh confirmation hearings in October, 2018 (Study 1). We choose this population because we expected these users' tweet histories to contain a sufficient amount of outrage to examine reinforcement learning effects. To test how results generalized to less politically engaged users, we also collected the same number of users (3669 tweet histories) who tweeted at least once about the United Airlines passenger mistreatment incident (Study 2). Across both studies we collected 7,331 users and 12.7 million total tweets. See **Materials and Methods** and **Fig. 1** for further details about data collection and validation of characteristics of the two samples. Data collection and analysis parameters were preregistered at https://osf.io/dsj6a (Study 1) and https://osf.io/nte3y (Study 2).

In each dataset, we ran time-lagged regression models to examine the association between the previous day's social feedback for outrage expressions and a given day's amount of outrage expression. We used generalized estimating equations (GEE) with robust standard errors (*44*) to estimate population-level effects treating tweets nested within users. Daily amounts of outrage tweets were modeled using a negative binomial distribution (*45*). Our main model estimated the effect of a previous day's outrage-specific feedback on the current day's outrage expression while statistically adjusting for the following variables: daily tweeting frequency; the users' number of followers; the presence of URLs or media in each tweet; the past week's amount of outrage expressions and outrage-specific feedback (to account for autocorrelation effects between past and present outrage expressions and the feedback those receive); and feedback that was not specific to outrage (to account for the fact that people tend to tweet everything more when they receive more feedback, and to demonstrate specificity in the effect of outrage-specific feedback on subsequent outrage expression). These model parameters were preregistered for both Study 1 and Study 2 (see **Materials and Methods**). We also show that results reported below are robust to models that treat time as a fixed and random factor, which measure how the population-average effect of social feedback changes over time, and account for variation in day-specific events ("exogenous shocks") that could impact outrage expression, respectively (see SOM, Section 2.0).

Supporting our hypotheses, we found that daily outrage expression was significantly and positively associated with the amount of social feedback received for the previous day's outrage expression (Study 1: $b = 0.03$, $p < .001$, 95% CI = [0.03, 0.03]; Study 2: $b = 0.02$, $p < .001$, 95% CI = [0.02, 0.03]). For our model, this effect size translates to an expected 2-3% increase in outrage expression on the following day of tweeting if a user received a 100% increase in feedback for expressing outrage on a given day. For instance, a user who averaged 5 likes/shares per tweet, and then received 10 likes/shares when they expressed outrage, would be expected to increase their outrage expression on the next day by 2-3%. While this effect size is small, it can easily scale on social media over time, become notable at scale at the network level, or for users who maintain a larger followership and could experience much higher than 100% increases in social feedback for tweeting outrage content (e.g., political leaders). For other model specifications to test the robustness of the effect, see SOM, Section 2.0.

A classical finding in the reinforcement learning literature is that reinforcement effects on behavior tend to diminish over time as the relationships between actions and outcomes are learned (*46*, *47*). Accordingly, we next considered the possibility that our model is underestimating the magnitude of the effect of social reinforcement on outrage expression because we are studying users who already have a long history of tweeting and receiving feedback (a minimum of 1 month up to many years of tweeting). Users with longer reinforcement histories may be less sensitive to recent feedback after larger earlier adjustments of their behavior. To test this possibility, we ran a

model where the length of users' learning histories (i.e., the more days they had tweeted and received feedback) was allowed to interact with the recent effects of social reinforcement. This model demonstrated a significant negative interaction between previous social feedback and days tweeted when predicting current outrage expression, indicating that the longer a users' reinforcement history, the smaller the effect of recent social feedback on outrage expression (Study 1: b = -0.02, p <.001, 95% CI = [-0.02, -0.01]; Study 2: b = -0.02, p <.001, 95% CI = [-0.03, -0.01].

Our observation that outrage expression on a given day increases in tandem with social feedback for the previous day's outrage expression is broadly consistent with the principles of reinforcement learning (*19*). However, reinforcement learning theory also suggests a more specific hypothesis: increases in current outrage expression should be related to previous outrage-specific social feedback that is higher or lower than expected, i.e., that generates a prediction error (*48*). To test this hypothesis, we created positive and negative prediction error variables by computing positive and negative differences between the mean of the previous 7 days' outrage-specific social feedback and the previous day's outrage-specific social feedback (see SOM, Section 2.3 for more details). This analysis revealed a significant, positive relationship between positive prediction errors from previous tweeting and current outrage expression in both studies. In this case, greater positive prediction errors on the previous day were associated with greater outrage expression on a given day, (Study 1: b = 0.01, p = <.001, 95% CI = [0.01, 0.02], Study 2: b = 0.02, p = <.001, 95% CI = [0.02, 0.03]). Meanwhile, negative prediction errors were negatively associated with outrage expression on the next day in Study 1 (b = -0.03, p = <.001, 95% CI = [-0.03, -0.02]). However, this effect was not replicated in Study 2 as there was no reliable effect of negative prediction error on subsequent outrage expression (b = 0.05, p = .325, 95% CI = [-0.04, 0.15]).

Above, we found that DOC shows discriminant validity against classifications of the broader category of negative sentiment. Here, we explored whether we observe similar effects of social reinforcement on negative sentiment expressions as we do for moral outrage expressions. Toward this end, we re-ran our main model replacing the outrage expression variable with a negative sentiment expression variable, as determined by the NSC described above. In this case, we conducted a conservative test by tuning the NSC so that its classifications of negative sentiment matched the distribution of negative sentiment extremity in tweets classified as outrage by DOC (see SI Appendix, Section 2.4). Thus, any differences observed cannot be explained by differences in sentiment extremity, but rather are from differences in the specificity of moral outrage relative to the broader category of negative sentiment. The dependent variable was a given day's negative sentiment expression and the main predictor variable was the lagged negative-sentiment-specific social feedback (see SI Appendix, Section 2.4). These models showed inconsistent results across datasets: in politically engaged users, we observed a significant positive effect of social reinforcement on subsequent negative sentiment expressions, albeit with a smaller effect size than was observed for moral outrage expressions in the same users (Study 1: b = 0.01, p <.001, 95% CI = [0.01, 0.01]). For less politically engaged users, however, the effect of social reinforcement on subsequent negative sentiment expressions was null (Study 2: b = -0.00, p = .338, 95% CI = [-0.01, 0.00]). These findings provide preliminary evidence that outrage expressions are more readily predicted by previous social feedback than expressions of negative sentiment more broadly.

Norm learning hypothesis

Next, we tested a hypothesis that norm learning processes impact online outrage expressions. We approached this question in two steps. First, we reasoned that in the context of the political topics we study here, outrage expressions should be more prevalent in social

networks populated by more ideologically extreme users. This logic is based on evidence that ideological extremity predicts outrage expression (*30, 49, 50*). More specifically, we predicted that individual users who are embedded within more ideologically extreme networks should be more likely to express outrage, *over and above* their own political ideology. In other words, if norm learning guides outrage expression, individual users should be more likely to express outrage when they are surrounded by others expressing outrage, regardless of their personal ideology.

To test this, we gathered data about the social network composition of the users in our datasets ('egos'), including the full list of users who follow each ego ('followers') and the full list of users followed by each ego ('friends'). This yielded a total of 6.28 million friends and followers for egos in the Kavanaugh dataset, and a total of 21 million friends and followers for egos in the United dataset. We used these data to estimate the ideological extremity of each ego in our dataset *(*51*)*, as well as all of each ego's friends and followers, yielding estimates of each ego's network-level ideological extremity (see **Fig. 1**).

As expected, we observed higher network-extremity in our politically engaged users (Kavanaugh dataset, Study 1) than in our less politically engaged users (United dataset, Study 2; **Fig. 1**). However, there was substantial variation in network-extremity in both datasets. We exploited this variability to test whether egos were more likely to express outrage in networks with more ideologically extreme members, statistically adjusting for users' own ideological extremity. We confirmed this was the case (Study 1: $b = 0.13$, $p < .001$, 95% CI = [0.10, 0.15]; Study 2, $b = 0.31$, $p < .001$, 95% CI = [0.26, 0.36]; **Fig. 1**). As can be seen in **Fig. 1**, network-extremity impacts outrage expression both between and within datasets: users in the Kavanaugh dataset, who on average are embedded in more extreme networks than users in the United dataset, show higher levels of outrage expression than users in the United dataset. In addition, within each dataset, users embedded within more extreme networks show higher levels of outrage expression.

Testing the difference between moral outrage expression and the broader category of negative sentiment, we found that users embedded within more ideologically extreme networks also expressed significantly more negative sentiment for Study 1, $b = 0.03$, $p < .001$, 95% CI = [0.01, 0.04] but not for Study 2, $b = -0.01$, $p = .679$, 95% CI = [-0.06, 0.04]. Furthermore, the effect of network-extremity in Study 1 showed a substantially weaker relationship with negative sentiment than with moral outrage (with the size of the negative sentiment effect being less than half the size of the moral outrage effect). This finding suggests that moral outrage expressions are more closely related to a social network's ideological extremity than voicing negative emotions more broadly. This is expected from a functionalist perspective of emotion expression, since moral outrage is more specifically tied to the domain of political ideology than the broader category of negative sentiment (*42*).

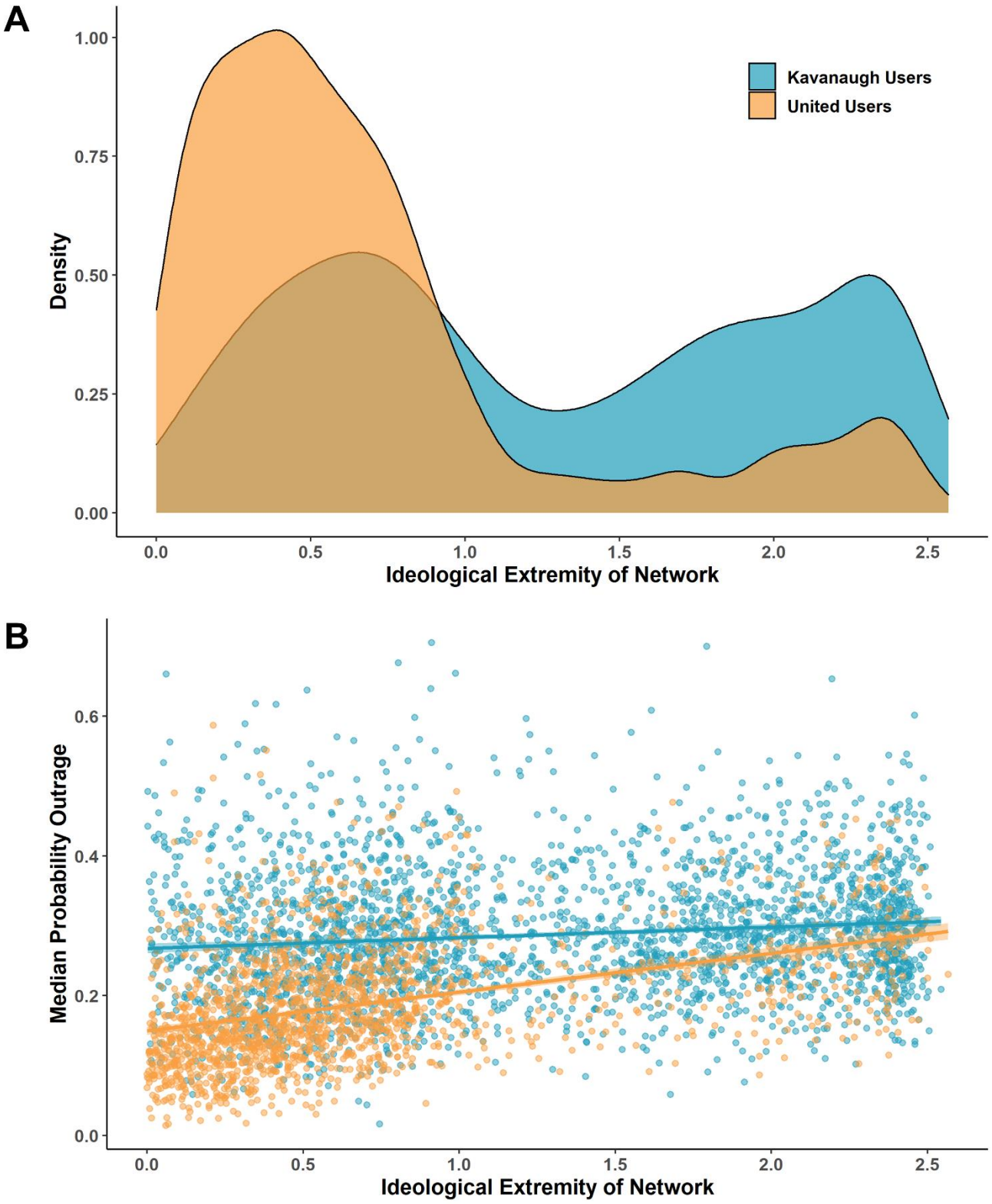

Fig. 1. Distributions of ideological extremity of user networks and levels of outrage expression. Panel A displays density plots of the ideological extremity of user networks for the Kavanaugh dataset (Study 1) and United dataset (Study 2). The x axis represents a continuous estimate of the mean ideological extremity of a user's network, greater values represent greater extremity. Panel B displays each user's median probability of expressing outrage in their tweets as a function of the ideological extremity of their network.

Second, we built on previous work demonstrating that individuals rely less on reinforcement learning to guide behavior when they are directly instructed which actions are valuable (*22*). One key prediction from recent theories of social learning is that information about social norms may be 'internalized' by learners (*21*), making them less responsive to local feedback from peers. Simply put, if a user can glean the appropriateness of outrage expression in their network by observing their newsfeed, they have less of a need to rely on reinforcement learning. This suggests that egos embedded in more ideologically extreme networks will be less sensitive to peer feedback in adjusting their outrage expressions.

To test this, we added ego-level and network-level ideological extremity as predictors to our time-lagged regression models examining social reinforcement of outrage, allowing both ego-extremity and network-extremity to interact with the social feedback effect. This analysis revealed that network-extremity significantly moderated the impact of social feedback on outrage expression, such that users embedded within more extreme networks showed weaker effects of social feedback on outrage expression (Study 1: b = -0.02, p = .004, 95% CI = [-0.03, -0.01]; Study 2: b = -0.05, p <.001, 95% CI = [-0.08, -0.02]), see **Fig. 2**. Meanwhile, ego-extremity did not moderate the impact of social feedback on outrage expression (Study 1: b = 0.01, p = .167, 95% CI = [0.00, 0.03]; Study 2: b = -0.02, p = .147, 95% CI = [-0.04, 0.01]). These results suggest that network-level norms of outrage expression moderate reinforcement learning over and above individual variation in ideology. More broadly, this finding supports the idea that to understand variation in users' outrage expression, it is important to consider both reinforcement learning and the frequency of outrage present in a network that users can observe to learn norms in their network. Users who infer that outrage is normative from its frequency in their network have less of a need to exclusively rely on reinforcement learning from social feedback to guide their outrage expressions. For negative sentiment expression, we found inconsistent results for the interaction of sentiment-specific feedback and network ideological extremity (Study 1: b = -0.01, p = .060, 95% CI = [-0.02, 0.00]; Study 2: b = -0.04, p = .018, 95% CI = [-0.08, -0.01]).

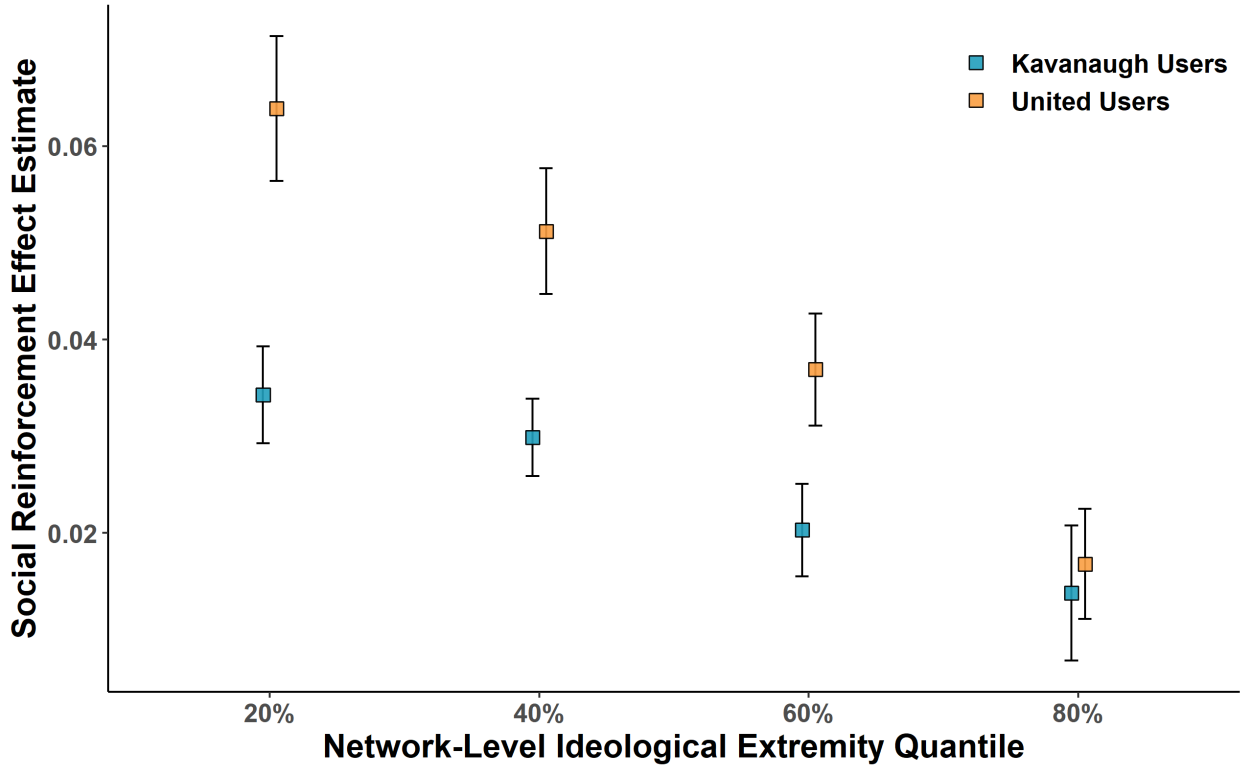

Fig. 2. Network-level ideological extremity moderates the effect of social feedback on outrage expressions. Each point displays the effect size estimate of previous social feedback predicting current outrage expression. Error bars were calculated based on standard errors of the estimate. The X axis represents quantile breaks from 20 to 80 percent. The blue color represents the Kavanaugh dataset users (Study 1), and the orange color represents the United dataset users (Study 2).

In summary, Studies 1 and 2 demonstrated three key findings: (1) outrage expression on Twitter can be explained in part by variation in social feedback that people receive via the platform; (2) users are more likely to express outrage in more ideologically extreme social networks; and (3) in more ideologically extreme social networks, users' outrage expression behavior is less sensitive to social feedback. These findings support our hypothesis that outrage expression on social media is shaped by both reinforcement learning and norm learning.

However, our observational approach in Studies 1 and 2 has several limitations. First, we cannot draw causal inferences about how social feedback or network-level norms shape outrage expressions, which limits the claims we can make about reinforcement learning and norm learning processes. Relatedly, we cannot rule out the possibility that social network composition might be endogenous to individuals' outrage expression. In other words, the effects we documented might also reflect the possibility that users who express more outrage may be more likely to follow more ideologically extreme users. This would suggest a different causal story than users learning to express outrage based on norms established by more extreme users. There is a high likelihood that both processes occur in tandem and feed into one another, as the joint influence of learning and self-selection into networks or social media platforms has been examined in recent work (*35*, *52*).

Finally, while we demonstrated a relationship between network-level ideological extremity and individual outrage expressions, it was computationally intractable to measure levels of outrage expression in the full tweet histories of >27 million users, which meant we could not directly measure network-level norms of outrage expression. We addressed these limitations with behavioral experiments in Studies 3 and 4.

Study 3

Study 3 directly manipulated social feedback and network-level norms of outrage expression in a simulated Twitter environment. The study was pre-registered at https://osf.io/rh2jk. Participants (N = 120) were randomly assigned to either an "outrage norm" or a "neutral norm" condition where they could scroll through a "newsfeed" containing 12 tweets from their "new" social network (**Fig. 3,** "Scrolling Stage"). Stimuli consisted of real tweets sampled from four contentious political topics, and outrage tweets were those classified as containing outrage expression by DOC (see **Materials and Methods**). In the outrage norm condition, 75% of the tweets contained outrage expressions and 25% contained neutral expressions. The outrage tweets displayed more likes and shares than the neutral tweets, in line with actual Twitter data (*29*, *30*). In the neutral norm condition, all tweets contained neutral expressions and displayed likes and shares in line with the 25% of neutral tweets displayed in the outrage norm condition. Participants were instructed to try and learn the content preferences of their new network (see SOM Appendix E for full instructions).

Participants then completed 30 trials of a learning task (**Fig. 3,** "Learning Stage") where they were incentivized to maximize social feedback (likes) from their network that were ostensibly provided by previous participants. On each trial, participants chose between two political tweets to "post" to the network (1 outrage, 1 neutral) and subsequently received feedback. Choosing an outrage tweet yielded greater social feedback on average. Our task design therefore allowed us to test the causal impact of social reinforcement on subsequent outrage expressions. Because the learning task was identical for participants in both the outrage norm and neutral norm conditions, we were also able to test the causal impact of norm information on subsequent reinforcement learning. We operationalized norm learning as a tendency to select the normative stimulus on the first trial of the learning task (outrage tweet in the outrage norm condition; neutral tweet in the neutral norm condition). We operationalized reinforcement learning as a tendency to increase selection of the positively reinforced stimulus over time (outrage tweets in both norm conditions).

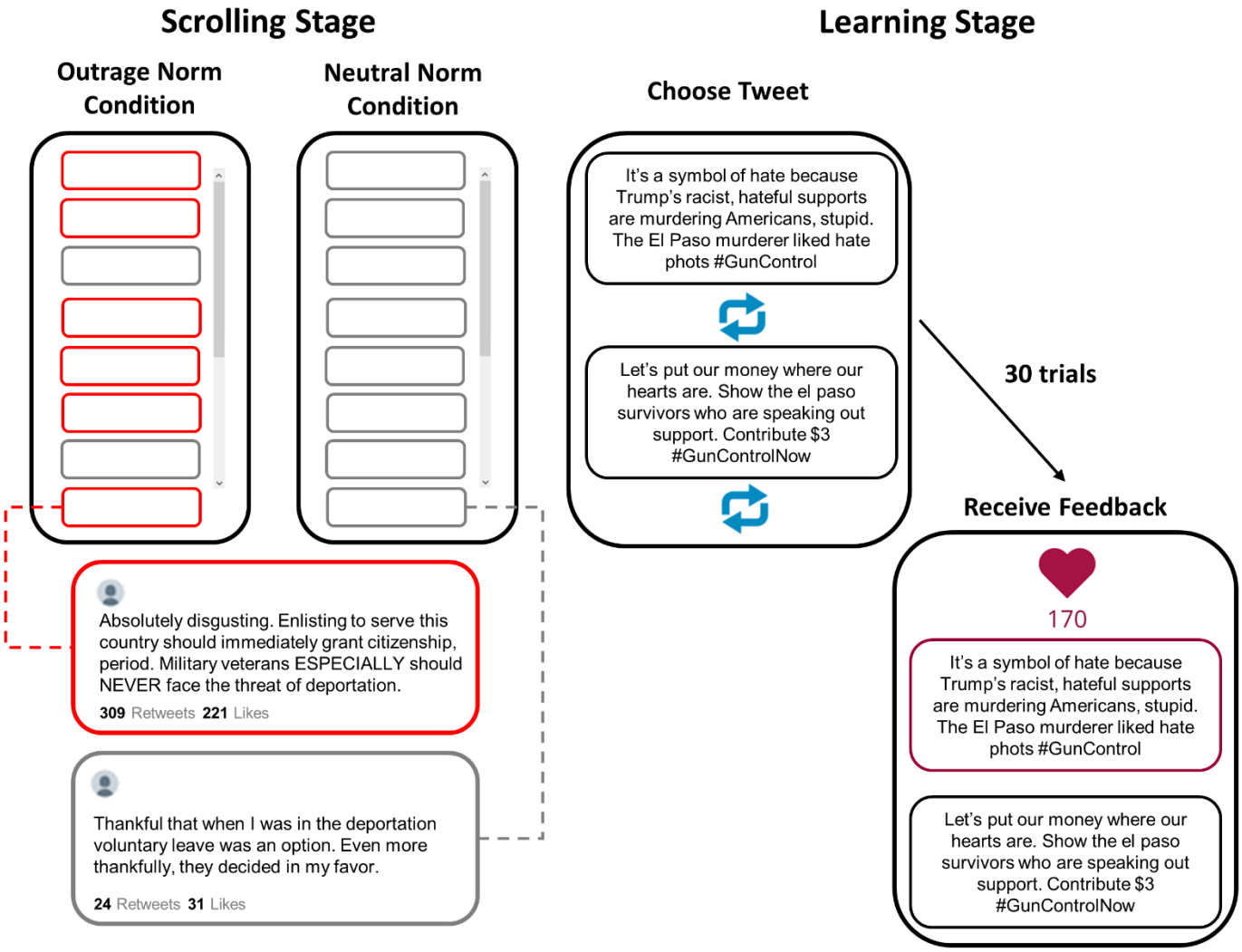

Fig. 3. Depiction of social media learning task (Studies 3 and 4). Participants first viewed what types of expressions were normative in their network by scrolling through 12 tweets. Next, they participated in a learning task where their goal was to maximize feedback.

Results confirmed that both reinforcement learning and norm learning shape outrage expression. As evidence of norm learning, on the first trial participants in the outrage norm condition were significantly more likely to select an outrage tweet than a neutral tweet, Odds Ratio *(OR)* = 4.94, $p < .001$, 95% CI = [3.10, 7.89], and participants in the neutral norm condition were significantly more likely to select a neutral tweet than an outrage tweet, *OR* = 1.73, $p < .001$, 95% CI = [1.11, 2.69]. In addition, we found evidence for reinforcement learning across both norm conditions, *OR* = 1.10, $p < .001$, 95% CI = [1.08, 1.12]. That is, participants learned to select more outrage tweets over time as a result of the trial-wise social feedback, see **Fig. 4A**. Simple effects revealed that participants in both the outrage norm condition (*OR* = 1.04, $p < .001$, 95% CI = [1.03, 1.08]) and the neutral norm condition (*OR* = 1.10, $p < .001$, 95% CI = [1.08, 1.12]) learned from social feedback to express more outrage over the course of the experiment.

However, the reinforcement learning effect was significantly smaller in the outrage norm condition than the neutral norm condition, as indicated by a significant negative interaction between the reinforcement learning effect and norm condition, *OR* = 0.95, $p < .001$, 95% CI = [0.92, 0.97], see **Fig. 4A**. This suggests that participants in the outrage norm condition relied on social feedback less to guide their outrage expressions, consistent with the findings of Studies 1

and 2.

Study 4

Study 4 (N = 120) replicated and extended Study 3 by testing whether the relative reliance on norm learning vs. reinforcement learning is similar for outrage expressions compared to neutral expressions. The study was pre-registered at https://osf.io/9he4n/. We used the same paradigm as in Study 3, with one critical difference: in the learning stage, participants received greater social feedback on average for the norm-congruent expression. That is, participants in the outrage norm condition received more positive feedback for selecting outrage tweets, while participants in the neutral norm condition received more positive feedback for selecting neutral tweets. This design allowed us to directly compare participants' reliance on norm learning versus reinforcement learning, for outrage expressions versus neutral expressions. As in Study 3, we operationalized norm learning as a tendency to select the normative stimulus on the first trial of the learning task (outrage tweet in the outrage norm condition; neutral tweet in the neutral norm condition). We operationalized reinforcement learning as a tendency to increase selection of the positively reinforced stimulus over time (outrage tweets in the outrage norm condition; neutral tweets in the neutral norm condition).

We again find evidence for norm learning: on the first trial participants in the outrage norm condition were more likely to select an outrage tweet than a neutral tweet, OR = 5.38, p < .001, 95% CI = [3.48, 8.31], while participants in the neutral norm condition were more likely to select a neutral tweet than an outrage tweet, OR = 1.54, p < .001, 95% CI = [1.03, 2.28]. We also find evidence for reinforcement learning: social feedback impacted participants' posting of outrage expressions, OR = 1.03, p < .001, 95% CI = [1.01, 1.05] as well as neutral expressions, OR = 1.06, p < .001, 95% CI = [1.05, 1.08]. Finally, we found that the reinforcement learning effect was smaller in the outrage norm condition compared to the neutral norm condition, as indicated by a significant interaction between the reinforcement learning effect and norm condition, Odds Ratio *(OR)* = 0.97, p < .001, 95% CI = [0.95, 0.99], see **Fig. 4**. In other words, participants in the outrage norm condition relied less on social feedback to guide their outrage expression compared to participants in the neutral norm condition.

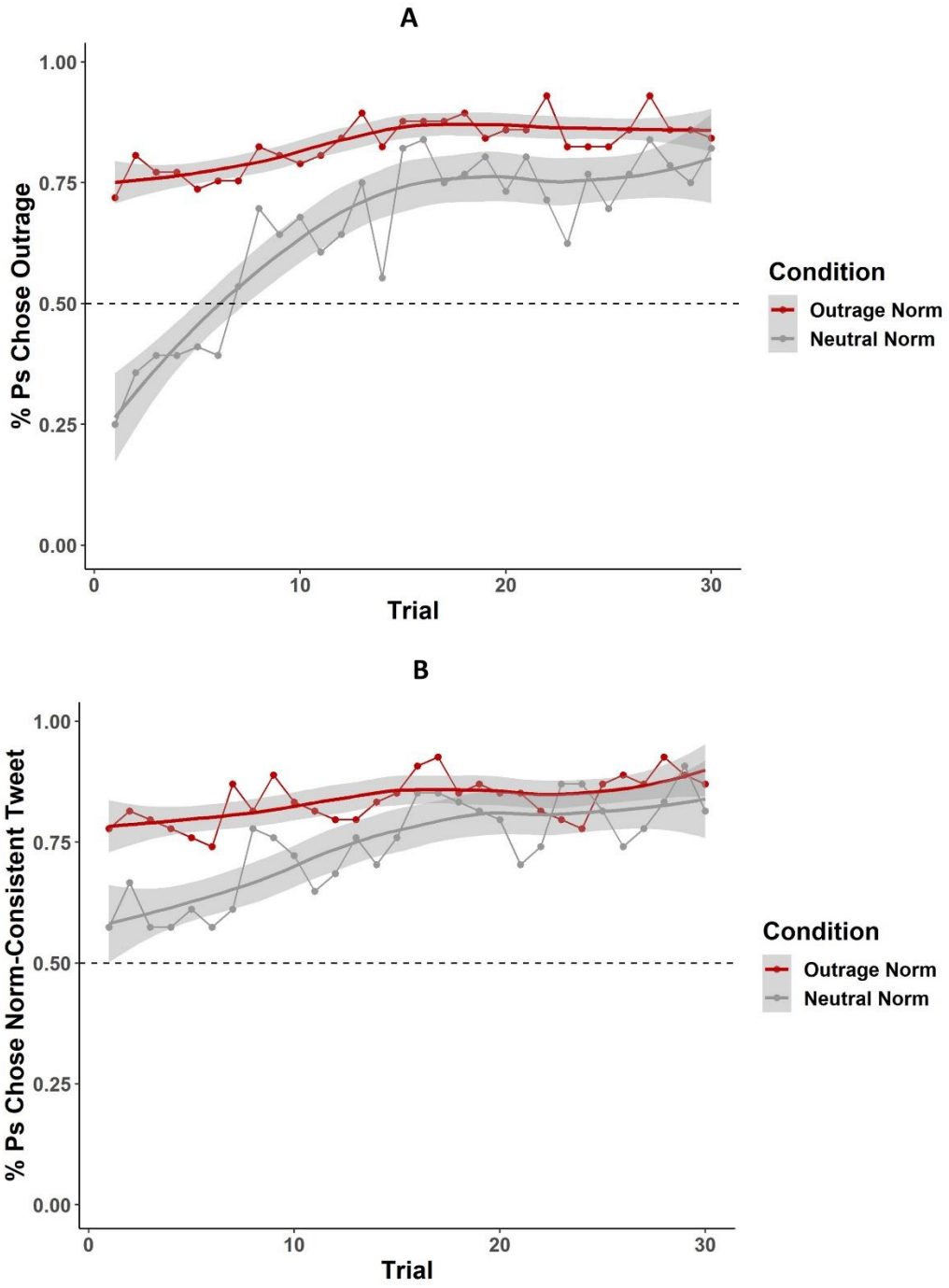

Fig. 4. Reinforcement learning and norm learning shape outrage expressions in a simulated social media environment. The y axis represents the percentage of participants on each trial that selected outrage tweets to post. The x axis represents the trial number. The red line represents participants in the outrage norm condition while the grey line represents participants in the neutral norm condition. Error bands represent the standard errors produced by fitting with a GAM function in *R* 3.6.1. The dotted line represents a 50% selection rate for participants in a given trial. Panel A displays results for Study 1, Panel B displays results for Study 2.

Discussion

Across two observational studies analyzing the tweet histories of 7,331 total users (12.7 million total tweets) and with two behavioral experiments (total $N = 240$), we investigated how reinforcement learning and norm learning shape moral outrage expressions on social media. Our findings revealed three key discoveries about moral outrage in the digital age. First, social feedback specific to moral outrage expression significantly predicts future outrage expressions, suggesting that reinforcement learning shapes users' online outrage expressions. Second, moral outrage expressions are sensitive to expressive norms in users' social networks, over and above users' own preferences, suggesting that norm learning processes guide moral expressions online. Third, network-level norms of expression moderate the social reinforcement of outrage: in networks that are more ideologically extreme, where outrage expression is more common, users are less sensitive to social feedback when deciding whether to express outrage. These findings underscore the importance of considering the interaction between human psychological tendencies and new affordances created by the specific design of social media platforms (*26*, *38*, *39*) to explain moral behavior in the digital age. This perspective dovetails with recent work in human-computer interaction research suggesting that consequential moral and political social media phenomena (e.g. the spread of disinformation) are best understood as a combination of top-down, orchestrated influence from powerful actors and bottom-up, participatory action from unwitting users (*35*, *53*).

At first blush, documenting the role of reinforcement learning in online outrage expressions may seem trivial. Of course, we should expect that a fundamental principle of human behavior, extensively observed in offline settings, will similarly describe behavior in online settings (*28*). However, reinforcement learning of *moral* behaviors online, combined with the *design* of social media platforms, may have especially important social implications. Social media newsfeed algorithms can directly impact how much social feedback a given post receives by determining how many other users are exposed to that post. Because we show here that social feedback impacts users' outrage expressions over time, this suggests newsfeed algorithms can influence users' moral behaviors by exploiting their natural tendencies for reinforcement learning. In this way, reinforcement learning on social media differs from reinforcement learning in other environments because crucial inputs to the learning process are shaped by corporate interests (*26*, *54*). Even if platform designers do not intend to amplify moral outrage, design choices aimed at satisfying other goals -- such as profit maximization via user engagement -- can indirectly impact moral behavior because outrage-provoking content draws high engagement (*29–31*). Given that moral outrage plays a critical role in collective action and social change (*42*, *55*), our data suggest that platform designers have the ability to influence the success or failure of social and political movements, as well as informational campaigns designed to influence users' moral and political attitudes (*35*, *53*) . Future research is required to understand whether users are aware of this, and whether making such knowledge salient can impact their online behavior.

Our findings also highlight other aspects of reinforcement learning that may be unique to the context of online social networks. First, we find consistent effects of positive prediction errors on reinforcement learning, but inconsistent effects of negative prediction errors. This may be due to the fact that social media platform design makes positive feedback ('likes' and 'shares') highly salient, while negative feedback (the absence of 'likes' and 'shares') is less salient. This design feature could make it much more difficult to learn from negative than positive feedback in online environments. Second, because users can self-organize into homophilic networks with easily observable communicative norms (*56*), following those norms might sometimes supersede reinforcement learning. We observe that in ideologically extreme networks where outrage expressions are more common, individual users are less sensitive to the social feedback they do receive, perhaps because the social feedback is redundant with information they gleaned from

observation, or because they have internalized network-level norms of expression (*21*). Crucially, our experimental data suggest that the context of social media makes the interaction of network norms and reinforcement learning especially likely to affect learning of expressions that convey reputational information to one's social group, like moral outrage (*57*). Future work may investigate how other properties of social networks impacts the balance between norm learning and reinforcement learning.

It is important to note that all of our conclusions concern the *expression* of moral outrage in social media text, and not the emotion itself, which we were unable to measure directly. Although in theory the experience and expression of moral outrage should be highly correlated, one intriguing possibility is that the design of social media platforms decouples expressions of outrage from experiencing the emotion itself (*13*, *26*). Such decoupling has implications for accounts of "outrage fatigue" – the notion that experiencing outrage is exhausting and thus diminishes over time. If expression becomes decoupled from experience, then outrage online may appear immune to fatigue even when experiencing it is not. Determining the extent to which expressions of emotion online represent actual emotional experiences is critical because if the social media environment decouples outrage expressions from experience, this could result in a form of pluralistic ignorance (*58*) whereby people falsely believe their peers are more outraged than they actually are (*26*).

This possibility is especially relevant in the context of political discourse (*59*, *60*), which has become increasingly polarized in recent years (*61*). Our findings may shed light on the rise of affective polarization -- intense, negative emotions felt toward political outgroups (*8*, *62*) that have erupted into violent clashes in the U.S. (*63*) and have been linked with inaccurate meta-perceptions of intergroup bias (*60*, *64*). In the current studies, we show that users conform to the expressive norms of their social network, expressing more outrage when they are embedded in ideologically extreme networks where outrage expressions are more widespread – regardless of their personal ideology. Such norm learning processes, combined with social reinforcement learning, might encourage more moderate users to become less moderate over time, as they are repeatedly reinforced by their peers for expressing outrage. Further studies that measure polarization longitudinally alongside social reinforcement and norm learning of outrage expressions will be required to test this prediction.

Our studies have several limitations. First, we note that all the users in our observational analyses were identified by having tweeted at least once during an episode of public outrage (though not all users necessarily expressed outrage during these episodes). This approach allowed us to ensure we collected a sample with a measurable signal of moral outrage, but it remains unclear whether these findings generalize to a broader population, other social media platforms, or outside the U.S. political context. Relatedly, Twitter users are not representative of the general population (*65*). However, they do comprise a high proportion of journalists and public figures who have an outsized influence on public affairs and the narratives surrounding them. Furthermore, our observational studies were unable to establish causal relationships between feedback, norms and outrage expression. We therefore chose to replicate the findings and demonstrate the causal relationship in tightly controlled experiments using mock social media environments (Studies 3 and 4). Although it would be scientifically interesting in future research to manipulate social feedback given to Twitter users, we caution that experimentally inducing changes in moral and political behavior in real online social networks raises a number of ethical concerns, especially considering that the majority of Twitter users are unaware their public data can be used for scientific study (*66*, *67*). An alternative possibility for future research is to recruit social media users who consent to participating in experiments where they are randomly assigned to conditions in which their social feedback experience is potentially modified.

There are also several limitations with our method for classification of moral outrage in social media text (DOC). As with all machine learning methods, DOC is not 100% accurate,

although we achieve performance on par with existing sentiment analysis methods that aim to classify more broad affective phenomena such as whether an expression is "positive" vs. "negative" (*68*). For this reason, within-sample estimates in changes of outrage over time might be more accurate than any single point-estimate for the purposes of generalizing out of sample. Furthermore, we note that we observed modest overlap between DOC's classifications of moral outrage and broader classifications of negative sentiment using existing classifiers (*69*), although social learning effects were stronger and more consistent for moral outrage expressions than negative sentiment. Although moral outrage is interesting to study due to its specific functional ties to morality and politics and the consequences it can bring about for individuals and organizations, more research is required to understand the extent to which our findings regarding moral outrage extend to other emotional expressions that are similarly tied to ideological extremity in politics such as fear (*70*). We also note that DOC is trained specifically on moral and political discourse in Twitter text, and therefore may have limited generalizability when applied to other social media platforms or other topics. As with all text classifiers, it is essential that researchers perform validity tests when applying DOC to a new sample before drawing conclusions from its results. Finally, we note that DOC was trained based on consensus judgments of tweets from trained annotators, which is useful for detecting broad linguistic features of outrage across individuals. However, specific social networks and even individuals may have diverse ways of expressing outrage, which suggests that future research should test whether incorporating individual-level or group-level contextual features can lead to greater accuracy in moral outrage classification (*71*).

Broadly, our results imply that social media platform design has the potential to amplify or diminish moral outrage expressions over time. Ultimately, whether it is "good" or "bad" to amplify moral outrage is a question that is beyond the scope of empirical studies, although leaders, policy-makers and social movements might assess whether online outrage achieves group-specific goals effectively (*6*, *72*). While our studies were not designed to assess the effectiveness of online outrage, we note that significant asymmetries have been documented along ideological and demographic lines, including the political right gaining far more political power from outrage in the media than the left (*73*), men gaining more status from anger than women (*74*), and anger mobilizing White people more than Black people in politics (*75*). These asymmetries might be exacerbated by social media platform design, in light of the growing impact of online discourse on political events and awareness (*76*, *77*). Future work is required to determine how online amplification of moral outrage might also spill over into offline social interactions, consumer decisions, and civic engagement.

Materials and Methods

Studies 1 and 2

Measuring moral outrage expressions in social media text. For our social media studies, we developed DOC using supervised machine learning. We trained DOC on a total of 26,000 tweets labeled as containing an expression of outrage or not, collected during a variety of episodes that sparked widespread public outrage (see **Table 1** for sources of training data and SOM, Section 1.0 for details of classifier development). Extensive evaluation demonstrated that DOC classified moral outrage expressions with accuracy and reliability comparable to extensively trained ('expert') human annotators (see SOM section 1.4 for details and **Table 2** for examples of tweets classified as containing moral outrage expression by DOC). DOC is available for academic researchers via a Python package at the following link: https://github.com/CrockettLab/outrage_classifier.

To develop DOC, we leveraged the Global Vectors for Word Representation (*78*) to encode tweets into vector space, and then input these word embeddings into a deep gated recurrent unit (*79*) neural network architecture (for tests of alternative models, see SOM, Section 1.0). The GRU model was trained on an initial data set of 16,000 tweets collected during a contentious political episode in American politics: the confirmation hearings of Supreme Court nominee Brett Kavanaugh (SOM, Section 1.1). Crucially, this episode sparked outrage from both liberals and conservatives, which made it ideal for training a classifier to detect aspects of outrage expressions that are not specific to a particular political ideology. We collected these tweets by gathering data on public mentions of politicians who were embroiled in controversy over statements about the confirmation hearings (see SOM, Section 1.1). We then trained 'crowdsourced' annotators to identify moral outrage expressions in these tweets based on social psychological theory (see SOM, Appendix A for full training instructions). Each tweet in the data set was rated as containing outrage or not by an ideologically heterogeneous group of 10 annotators (5 liberals and 5 conservatives). Annotators demonstrated excellent reliability in applying our criteria for identifying moral outrage expressions as assessed by an intraclass correlation: ICC(3,10) = .82, 95% CI = [0.82, 0.83]. Importantly, we found that when holding the number of annotators constant at 5, politically heterogenous groups (ICC(3,5) = .69) showed similar reliability as politically homogenous groups (mean ICC(3,5) = .70), justifying the combined use of liberal and conservative annotators to determine outrage ratings (for more details see SOM, Section 1.2).

We then collected a secondary set of various political topics and had them labeled by expert human annotators (N = 10,000) to order to enhance the domain-generalizability of DOC. We selected these topics to represent diverse moral transgressions that violated both liberal and conservative values, as well as a non-political moral transgression (see **Table 2** and SOM, Section 1.5). To test DOC's performance, we trained and tested on the 26,000-tweet labeled data set using 5-fold cross-validation and found that our GRU model achieved an accuracy of 75% and F-1 score of .71 in classification of moral outrage (see SOM, Section 1.0 for more details). Importantly, DOC applied outrage labels similarly to the expert annotators in a sample of 500 tweets: the reliability applying outrage labels for the group of 8 expert annotators (ICC(2,8) = .88, 95% CI = [.86, .89]), was statistically indistinguishable from the mean reliability of all possible groups comprising 7 expert annotators and DOC (ICC(2,8) = .87, 95% CI = [.86, .89]). In short, DOC classified moral outrage in a manner consistent with expert human annotators.

As moral outrage is a specific type of negative sentiment, we expected outrage expression and negative sentiment to be correlated, but not identical. Supporting this prediction, DOC showed discriminant validity comparing its classifications to the classifications of a model trained to identify the broader category of negative sentiment. When examining the classifications made by the two models in the 26,000-tweet labeled dataset, we observed a weak correlation, $\tau = .11$, $p < .001$. Descriptively, we observed that outrage and negative sentiment classifications showed agreement in only 29% of cases. See SOM Section 1.7 and Table S15 for more details and examples of tweets containing negative sentiment but not moral outrage expression.

Topic	Text	Classification
Kavanaugh	@SenGillibrand you are a DISGRACE. Shut your lying mouth. There is no evidence of assault	Outrage

Kavanaugh	@JeffFlake thank you for stepping up. Don't let them do a poor job in the investigation	Non-Outrage
Covington	I cannot with the "Stand with Covington" gofundme? WTF? People are giving these brats money? Unbelievable!	Outrage
Covington	There are good people on both sides of the #Covington debate. Let's all slow down	Non-Outrage
United	I'm in total disgust and madness because of what #united did. Totally Unacceptable.	Outrage
United	Here's the latest ad from @united. #united #advertising https://...	Non-Outrage
Smollett	Hey @JussieSmolett you are a worthless piece of shit. A greedy, corrupt liar.	Outrage
Smollett	We need some more @JussieSmolett memes. Where are they?	Non-Outrage
Transgender Ban	This is a disgusting display of hatred and oppression. #FUCKYOUTRUMP and your criminal cabinet!	Outrage
Transgender Ban	Hillary Clinton said some thoughtful words about the ban: https://...	Non-Outrage

Table 2. Example outrage and non-outrage tweets as classified by the Digital Outrage Classifier. The table shows example tweets from five political topics appearing in our training set that were classified as containing outrage vs. not containing outrage by DOC. To protect Twitter user privacy from reverse text searches, for figure display purposes only some words from each original tweet have been edited while maintaining salient features of the message.

Hypothesis testing. To test our research questions regarding the social learning of outrage expressions, we first used metadata from our training dataset to select a group of Twitter users who were identifiable as authors of tweets in the Kavanaugh dataset, and who maintained public profiles for at least 3 months after the original data collection (3,669 users). We connected to Twitter's standard and premium APIs, and collected these users' full tweet histories yielding 6.1 million tweets available for analysis (see SOM, Section 2.0 for more details). We used the same method to collect a second group of less politically engaged users, who were identified as authors of tweets in the United Airlines dataset (3,669 Twitter users with 6.6 million tweets available for analysis). Since tweets in the United dataset did not concern a politically partisan issue, we expected that users identified from this dataset would be less ideologically extreme than the Kavanaugh users. Estimating the ideology of users in both the Kavanaugh and United datasets confirmed this (see SOM, Section 2.2). This analysis strategy enabled us to test to what extent our findings generalize across different levels of political engagement and ideological extremity.

To test the association between outrage and previously received social feedback, we used generalized estimating equations (*44*)) with robust standard errors (observations, or tweets, were

clustered by user) to estimate the population-level association between moral outrage expression and the amount of social feedback received on the previous day, with data aggregated at the level of days. We modeled the sum of outrage expression as a negative binomial distribution with a log link function and an independent correlation structure using PROC GENMOD in *SAS* 9.4. Decisions for modeling the outcome variable and correlation structure were based on the fact the outcome variable was overdispersed count data, and also on QIC model fit statistics (*80*) available in PROC GENMOD. To replicate the analyses in *R* 3.6.1 in a computationally efficient manner, we used the 'bam' function in the package 'mgcv' v1.8. *SAS* and *R* scripts used for data organization and model estimation described in this section are available at: https://osf.io/9he4n/. Model specifications and variable formations listed below were pre-registered at https://osf.io/dsj6a (Study 1) and https://osf.io/nte3y (Study 2).

The model predicting outrage expression from previous social feedback included as predictors the sum of feedback received when outrage was expressed for 7 lagged days, previous outrage tweeting for 7 lagged days, previous sums of non-outrage feedback for 7 lagged days, user-level tweet history total, number of tweets containing URLS per day, number of tweets containing media per day, and the user follower count. Results were robust to various model specifications including a model that included one 1 previous day of outrage feedback, previous tweeting, and feedback for non-outrage tweets (i.e., including only 1 lag for each variable). Results were also robust when modeling the main lagged predictor variable as the difference between feedback received for outrage tweets vs. non-outrage tweets (i.e., what is the effect of receiving more feedback for outrage expression compared to other tweets a user sent?). SOM Section 2.0 presents full model details and tabulated results.

We created positive and negative prediction error variables by computing the difference between the previous day's outrage-specific social feedback and the feedback from 7 days previous to the first lag. For example, if a user received an average of 5 likes/shares across days t-2 – t-8, and on day t-1 they received 8 likes and shares, that observation would be recorded as a +3 for the positive prediction error variable and a 0 for the negative prediction error variable. If on day t-1 they received 3 likes and shares, the observation would be recorded as a -2 for the negative prediction variable and a 0 for the positive prediction error variable. Further details are presented in SOM, Section 2.3.

To test norm learning hypotheses in the Kavanaugh and United Airlines datasets, we defined the social network of each 'ego' (a user in a dataset) as all friends and followers of the ego, and estimated the political ideology of each user in the ego's network (*51*). We defined ideological extremity as the absolute value of the mean political ideology of all users in an ego's social network (thus, higher values represent more extreme users, see SOM, Section 2.0 for more details). To test how network ideological extremity moderated the social reinforcement effects, we regressed daily outrage expression on the two-way interaction of the previous day's outrage-specific feedback and each ego's network ideological extremity while also adjusting for daily tweet frequency and covariates included in above models. SOM Section 2.0 presents full model details and tabulated results.

Study 3

Participants. We recruited 120 participants via the Prolific participant recruitment platform. We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study in our preregistration at https://osf.io/rh2jk. Participants were all liberal as our Twitter stimuli express left-leaning opinions about contentious political topics.

Procedure. Participants were recruited to participate in a simulated Twitter environment and told they were a new member of an ostensible network of platform users. They were explicitly instructed to learn the content preferences of their "new" network (For full instructions see SOM, Section 3.0. Participants were randomly assigned to either an 'outrage norm' or 'neutral norm' condition. Both conditions consisted of two stages: a scrolling stage and a learning stage (**Fig. 3Fig. 3**). In the scrolling stage, participants passively viewed 12 tweets that were sent from their new network by scrolling through a simulated Twitter "newsfeed". Each tweet commented on one of four contentious political topics: (1) the first impeachment of Donald Trump as US president, (2) Medicare for All, (3) US immigration policy, and (4) the 'extinction rebellion' climate change movement. Each tweet discussed one of these issues from a liberal perspective. Three tweets from each of the topics were selected and combined to make the 12 tweets participants viewed.

The tweet stimuli were collected from publicly available tweets (no usernames were displayed for the tweet stimuli), and outrage expression was determined using DOC. In the outrage norm condition, 75% of the tweets that participants saw contained an expression of outrage, while the remaining 25% did not. None of the tweets seen by participants in the neutral norm condition contained outrage. Whether a tweet contained outrage or not was determined by using DOC to classify the tweets and then checking for validity of classification.

In addition to manipulating the prevalence of outrage in each condition, the amount of positive social feedback (i.e., 'likes') displayed under each tweet was also varied. In the outrage norm condition, tweets that contained expressions of outrage displayed an amount of likes randomly drawn from a 'high reward distribution' (M = 250, SD = 50). Non-outrage tweets in this condition were assigned a number of likes sampled from a much lower distribution (M = 25, SD = 6). In the neutral norm condition, a random selection of 75% of the tweets in the neutral condition had high feedback, 25% had low feedback as determined by the same distributions in the outrage norm condition.

After completing the scrolling stage, participants completed a learning stage where their goal was to maximize the social feedback they received for 'retweeting' content (i.e., re-posting a tweet to their network). Participants were incentivized to maximize their feedback via potential bonus payment related to total likes accumulated during the experiment. Social feedback was operationalized as Twitter 'likes', also known as 'favorites', which were ostensibly awarded by participants who previously completed the task and who shared the views of the network. On each of 30 trials, participants were presented with two new tweets discussing the same political topics that were used in the scrolling stage. As before, these tweets were classified for outrage expression using DOC. Thus, while both tweets in a pair discussed the same topic, one tweet contained outrage while the other did not. The position of the tweets when presented (left or right side of the screen) was randomized. Participants responded on each trial by clicking a 'retweet button' that corresponded to the member of the pair of tweets they wished to share. Once they clicked the retweet button, participants were immediately presented with the feedback awarded to the selected tweet.

The social feedback awarded on each trial was drawn from either of two predetermined 'reward trajectories' with the trajectory used determined by the participants retweet choice. For example, if a participant chose to retweet the outraged content in the n^{th} trail, then the feedback they were awarded corresponded to the n^{th} integer in an array of values. Of these values, 80% were randomly drawn from the high reward distribution used in the scroll task. The remaining 20% of reward values were sampled from the low distribution. These reward contingencies were the same for all participants, irrespective of the norm condition they were assigned in the scrolling task. The 80/20 split was used to add noise to the feedback and thus make it more difficult for participants to quickly infer the underlying reward structure.

Data Analysis. We modeled participants' binary tweet choices over trials using a generalized linear mixed model with the *'lme4'* package in *R 3.6.1*. Norm condition, trial number and their interaction were entered as fixed effects, and we entered a random intercept for participants. Results were robust to modeling stimulus as a random factor (*81*), see SOM, Section 3.0.

Study 4

Participants. We recruited 120 participants via the Prolific participant recruitment platform. We report how we determined our sample size, all data exclusions (if any), all manipulations, and all measures in the study in our preregistration at https://osf.io/jc9tq. Participants were all liberal as our Twitter stimuli express left-leaning opinions about contentious political topics.

Procedure. As in Study 3, participants completed a simulated Twitter task with a scrolling stage and a learning stage (**Fig. 3**). The scrolling stage was identical to that in Study 3. The learning stage was similar to that in Study 3, with one exception: participants in the neutral norm condition received more likes for selecting neutral tweets, while participants in the outrage norm condition received more likes for selecting outrage tweets. This design allowed us to directly compare learning rates in environments where outrage versus neutral tweets receive more positive feedback.

Data Analysis. We modeled participants' binary tweet choices over trials using a generalized linear mixed model with the *'lme4'* package in *R 3.6.1*. Norm condition, trial number and their interaction were entered as fixed effects, and we entered a random intercept for participants. Results were robust to modeling stimulus as a random factor (*81*), see SOM, Section 3.0

References and Notes

1. H. Gintis, S. Bowles, R. Boyd, E. Fehr, *Moral sentiments and material interests: The foundations of cooperation in economic life* (MIT Press, Cambridge, MA, 2005).

2. J. M. Salerno, L. C. Peter-Hagene, The interactive effect of anger and disgust on moral outrage and judgments. *Psychological Science*. **24**, 2069–2078 (2013).

3. P. E. Tetlock, O. V Kristel, S. B. Elson, M. C. Green, J. S. Lerner, The psychology of the unthinkable: taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of personality and social psychology*. **78**, 853–870 (2000).

4. E. Fehr, U. Fischbacher, Third-party punishment and social norms. *Evolution and Human Behavior*. **25**, 63–87 (2004).

5. B. Simpson, R. Willer, A. Harrell, The Enforcement of Moral Boundaries Promotes Cooperation and Prosocial Behavior in Groups. *Scientific Reports*. **7** (2017), doi:10.1038/srep42844.

6. V. Spring, D. Cameron, M. Cikara, The upside of outrage. *Trends in Cognitive Sciences* (2018).

7. D. G. Young, *Irony and outrage: The polarized landscape of rage, fear, and laughter in the United States* (Oxford University Press, USA, 2019).

8. E. J. Finkel, C. A. Bail, M. Cikara, P. H. Ditto, S. Iyengar, S. Klar, L. Mason, M. C. McGrath, B. Nyhan, D. G. Rand, L. J. Skitka, J. A. Tucker, J. J. Van Bavel, C. S. Wang, J. N. Druckman, Political sectarianism in America: A poisonous cocktail of othering, aversion, and moralization poses a threat to democracy. *Science*. **370**, 533–536 (2020).

9. J. Shepard, K. B. Culver, Culture Wars on Campus: Academic Freedom, the First Amendment and Partisan Outrage in Polarized Times. *San Diego Law Review*. **55**, 87–158 (2018).

10. M. Lynch, Do we really understand "fake news"? *The New York Times* (2019), (available at https://www.nytimes.com/2019/09/23/opinion/fake-news.html).

11. J. Haidt, T. Rose-Stockwell, The dark psychology of social networks. *The Atlantic* (2019), (available at https://www.theatlantic.com/magazine/archive/2019/12/social-media-democracy/600763/).

12. H. Gintis, *Moral sentiments and material interests: The foundations of cooperation in economic life* (MIT Press, Cambridge, Vol. 6., 2005).

13. M. J. Crockett, Moral outrage in the digital age. *Nature Human Behaviour*. **1**, 769–771 (2017).

14. A. Lieberman, J. Schroeder, Two social lives: How differences between online and offline interaction influence social outcomes. *Current Opinion in Psychology*. **31** (2020), pp. 16–21.

15. M. L. Hoffman, *Empathy and Moral Development* (Cambridge University Press, New York, NY, 2000).

16. L. Montada, A. Schneider, Justice and emotional reactions to the disadvantaged. *Social Justice Research*. **3**, 313–344 (1989).

17. J. Haidt, in *Handbook of affective sciences*, R. J. Davidson, K. R. Scherer, H. H. Goldsmith, Eds. (Oxford University Press, New York, NY, 2003), pp. 572–595.

18. J. M. Darley, Morality in the Law: The psychological foundations of citizens' desires to punish transgressions. *Annual Review of Law and Social Science*. **5**, 1–23 (2009).

19. J. Gläscher, N. Daw, P. Dayan, J. P. O'Doherty, States versus rewards: Dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron*. **66**, 585–595 (2010).

20. P. W. Glimcher, Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences of the United States of America*. **108**, 15647–15654 (2011).

21. M. K. Ho, J. MacGlashan, M. L. Littman, F. Cushman, Social is special: A normative framework for teaching with and learning from evaluative feedback. *Cognition*. **167**, 91–106 (2017).

22. J. Li, M. R. Delgado, E. A. Phelps, How instructed knowledge modulates the neural systems of reward learning. *Proceedings of the National Academy of Sciences of the United States of America*. **108**, 55–60 (2011).

23. R. B. Cialdini, C. A. Kallgren, R. R. Reno, A Focus Theory of Normative Conduct: A Theoretical Refinement and Reevaluation of the Role of Norms in Human Behavior. *Advances in Experimental Social Psychology*. **24**, 201–234 (1991).

24. N. Velez, H. Gwon, Learning from other minds: An optimistic critique of reinforcement learning models of social learning (2020), doi:https://doi.org/10.31234/osf.io/q4bxr.

25. B. T. R. Savarimuthu, R. Arulanandam, M. Purvis, in *Lecture Notes in Computer Science* (2011), pp. 36–50.

26. W. J. Brady, M. J. Crockett, J. J. Van Bavel, The MAD Model of Moral Contagion: The role of motivation, attention and design in the spread of moralized content online. *Perspectives on Psychological Sciences* (2020), doi:10.31234/osf.io/pz9g6.

27. M. Dehghani, K. Johnson, J. Hoover, E. Sagi, J. Garten, N. J. Parmar, S. Vaisey, R. Iliev, J. Graham, Purity Homophily in Social Networks. *Journal of Experimental Psychology: General*. **145**, 366–375 (2016).

28. B. Lindström, M. Bellander, A. Chang, P. Tobeler, D. M. Amodio, A computational reinforcement learning account of social media engagement. *PsyArXiv* (2019).

29. W. J. Brady, J. A. Wills, J. T. Jost, J. A. Tucker, J. J. Van Bavel, Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*. **114**, 7313–7318 (2017).

30. W. J. Brady, J. A. Wills, D. Burkart, J. T. Jost, J. J. Van Bavel, An Ideological Asymmetry in the Diffusion of Moralized Content on Social Media Among Political Leaders. *Journal of Experimental Psychology: General*. **148**, 1802–1813 (2019).

31. S. Valenzuela, M. Piña, J. Ramírez, Behavioral Effects of Framing on Social Media Users: How Conflict, Economic, Human Interest, and Morality Frames Drive News Sharing. *Journal of Communication* (2017), doi:10.1111/jcom.12325.

32. R. Cialdini, M. Trost, Social influence: Social norms, conformity and compliance. *The Handbook of Social Psychology, Vol. 2* (1998), pp. 151–192.

33. R. B. Cialdini, C. A. Kallgren, R. R. Reno, A Focus Theory of Normative Conduct: A Theoretical Refinement and Reevaluation of the Role of Norms in Human Behavior. *Advances in Experimental Social Psychology*. **24**, 201–234 (1991).

34. H. a Chapman, D. a Kim, J. M. Susskind, a K. Anderson, In bad taste: evidence for the oral origins of moral disgust. *Science (New York, N.Y.)*. **323**, 1222–1226 (2009).

35. K. Starbird, A. Arif, T. Wilson, Disinformation as collaborative work: Surfacing the participatory nature of strategic information operations. *Proceedings of the ACM on Human-Computer Interaction*. **3**, 1–26 (2019).

36. Y. Nagar, in *Proceedings of the ACM 2012 conference on computer supported cooperative work* (2012), pp. 393–402.

37. L. Y. Atlas, How instructions shape aversive learning: higher order knowledge, reversal learning, and the role of the amygdala. *Current Opinion in Behavioral Sciences*. **26** (2019), pp. 121–129.

38. S. K. Evans, K. E. Pearce, J. Vitak, J. W. Treem, Explicating Affordances: A Conceptual Framework for Understanding Affordances in Communication Research. *Journal of Computer-Mediated Communication*. **22**, 35–52 (2017).

39. J. B. Bayer, P. Triệu, N. B. Ellison, Social Media Elements, Ecologies, and Effects. *Annual review of psychology*. **71** (2020), doi:10.1146/annurev-psych-010419-050944.

40. J. Ronson, *So you've been publicly shamed* (Riverhead Books, New York, NY, 2016).

41. P. E. Tetlock, O. V. Kristel, S. B. Elson, M. C. Green, J. S. Lerner, The psychology of the unthinkable: Taboo trade-offs, forbidden base rates, and heretical counterfactuals. *Journal of Personality and Social Psychology*. **78**, 853–870 (2000).

42. C. Hutcherson, J. Gross, The moral emotions: A social–functionalist account of anger, disgust, and contempt. *Journal of personality and social …*. **100**, 719–737 (2011).

43. A. Go, R. Bhayani, Huang, Sentiment140 - A Twitter Sentiment Analysis Tool, (available at http://help.sentiment140.com/home).

44. J. W. Hardin, B. Everitt, D. Howell, Eds. (Wiley Online Library, Hoboken, NJ, 2005).

45. J. M. Hilbe, Negative Binomial Regression. *Public Administration Review*. **70**, 1–6 (2011).

46. A. Dickinson, Actions and Habits: The Development of Behavioural Autonomy. *Philosophical Transactions of the Royal Society B: Biological Sciences*. **308**, 67–78 (1985).

47. J. E. Grusec, J. J. Goodnow, Impact of Parental Discipline Methods on the Child's Internalization of Values: A Reconceptualization of Current Points of View. *Developmental Psychology*. **30**, 4–19 (1994).

48. M. Pessiglione, B. Seymour, G. Flandin, R. J. Dolan, C. D. Frith, Dopamine-dependent prediction errors underpin reward-seeking behaviour in humans. *Nature*. **442**, 1042–5 (2006).

49. G. E. Marcus, N. A. Valentino, P. Vasilopoulos, M. Foucault, Applying the Theory of Affective Intelligence to Support for Authoritarian Policies and Parties. *Political Psychology*. **40**, 109–139 (2019).

50. Partisan Conflict and Congressional Outreach. *Pew Research Center* (2017), (available at https://www.people-press.org/2017/02/23/partisan-conflict-and-congressional-outreach/).

51. P. Barberá, Birds of the same feather tweet together: Bayesian ideal point estimation using twitter data. *Political Analysis*. **23**, 76–91 (2015).

52. K. Munger, J. Phillips, A Supply and Demand Framework for YouTube Politics (2019).

53. A. Arif, L. G. Stewart, K. Starbird, Acting the part: Examining information operations within #BlackLivesMatter discourse. *Proceedings of the ACM on Human-Computer Interaction*. **2**, 1–27 (2018).

54. A. Goldenberg, J. J. Gross, Digital Emotion Contagion. *Trends in Cognitive Sciences*. **24**, 316–328 (2020).

55. M. Reifen Tagar, C. M. Federico, E. Halperin, The positive effect of negative emotions in protracted conflict: The case of anger. *Journal of Experimental Social Psychology*. **47**, 157–164 (2011).

56. S. Yardi, D. Boyd, Dynamic Debates: An Analysis of Group Polarization Over Time on Twitter. *Bulletin of Science, Technology & Society*. **30**, 316–327 (2010).

57. J. J. Jordan, D. G. Rand, Signaling When No One Is Watching: A Reputation Heuristics Account of Outrage and Punishment In One- Shot Anonymous Interactions. *Journal of Personality and Social Psychology* (2019), doi:10.1037/pspi0000186.

58. D. A. Prentice, D. T. Miller, Pluralistic Ignorance and Alcohol Use on Campus: Some Consequences of Misperceiving the Social Norm. *Journal of Personality and Social Psychology*. **64**, 243–256 (1993).

59. H. Hwang, Y. Kim, C. U. Huh, Seeing is Believing: Effects of Uncivil Online Debate on Political Polarization and Expectations of Deliberation. *Journal of Broadcasting and Electronic Media*. **58**, 621–633 (2014).

60. J. Lees, M. Cikara, Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts. *Nature Human Behaviour*. **4**, 279–286 (2020).

61. N. McCarty, K. Poole, H. Rosenthal, *Polarized America: The Dance of Ideology and Unequal Riches* (MIT Press, Cambridge, ed. 2nd, 2016).

62. J. C. Rogowski, J. L. Sutherland, How Ideology Fuels Affective Polarization. *Political Behavior*. **38**, 485–508 (2016).

63. S. Stolberg, B. Rosenthal, Man Charged After White Nationalist Rally in Charlottesville in Deadly Violence. *NY Times* (2017), (available at https://www.nytimes.com/2017/08/12/us/charlottesville-protest-white-nationalist.html).

64. S. L. Moore-Berg, L.-O. Ankori-Karlinsky, B. Hameiri, E. Bruneau, Exaggerated meta-perceptions predict intergroup hostility between American political partisans. *Proceedings of the National Academy of Sciences*. **117**, 14864–14872 (2020).

65. M. Duggan, J. Brenner, The Demographics of Social Media Users — 2012. *Pew Research Center2* (2012), (available at https://www.pewresearch.org/internet/2013/02/14/the-demographics-of-social-media-users-2012/).

66. C. Fiesler, N. Proferes, "Participant" Perceptions of Twitter Research Ethics. *Social Media and Society*. **4** (2018), doi:10.1177/2056305118763366.

67. B. Hallinan, J. R. Brubaker, C. Fiesler, Unexpected expectations: Public reaction to the Facebook emotional contagion study. *New Media and Society*. **22**, 1076–1094 (2020).

68. J. Barnes, R. Klinger, S. Schulte im Walde, in *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (Association for Computational Linguistics, Copenhagen, Denmark, 2017), pp. 2–12.

69. J. Silge, D. Robinson, tidytext: Text Mining and Analysis Using Tidy Data Principles in R. *Journal of Open Source Software*. **1**, 37 (2016).

70. J.-W. van Prooijen, A. P. M. Krouwel, M. Boiten, L. Eendebak, Fear Among the Extremes: How Political Ideology Predicts Negative Emotions and Outgroup Derogation. *Pers Soc Psychol Bull*. **41**, 485–497 (2015).

71. J. Garten, B. Kennedy, K. Sagae, M. Dehghani, Measuring the importance of context when modeling language comprehension. *Behav Res*. **51**, 480–492 (2019).

72. W. J. Brady, M. J. Crockett, How effective is online outrage? *Trends in Cognitive Sciences* (2018).

73. J. Hacker, P. Pierson, *Let them eat tweets: How the right rules in an age of extreme inequality* (Liveright, New York, NY, 2020).

74. V. L. Brescoll, E. L. Uhlmann, Can an angry woman get ahead? Status conferral, gender, and expression of emotion in the workplace: Research article. *Psychological Science* (2008), doi:10.1111/j.1467-9280.2008.02079.x.

75. D. L. Phoenix, *The Anger Gap: How Race Shapes Emotion in Politics* (Cambridge University Press, 2019).

76. J. T. Jost, P. Barberá, R. Bonneau, M. Langer, M. Metzger, J. Nagler, J. Sterling, J. A. Tucker, How Social Media Facilitates Political Protest: Information, Motivation, and Social Networks. *Political Psychology*. **39**, 85–118 (2018).

77. D. Freelon, A. Marwick, D. Kreiss, False equivalencies: Online activism from left to right. *Science*. **369**, 1197–1201 (2020).

78. J. Pennington, R. Socher, C. Manning, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)* (2014).

79. K. Cho, B. Van Merriënboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, Learning phrase representations using RNN encoder-decoder for statistical machine translation. *arXiv preprint arXiv:1406.1078* (2014).

80. W. Pan, Akaike's Information Criterion in Generalized Estimating Equations. *Biometrics*. **57**, 120–125 (2001).

81. C. M. Judd, J. Westfall, D. A. Kenny, Treating stimuli as a random factor in social psychology: A new and comprehensive solution to a pervasive but largely ignored problem. *Journal of Personality and Social Psychology*. **103**, 54–69 (2012).

82. S. Tatum, Brett Kavanaugh's nomination: A timeline. *CNN* (2018).

83. J. Roesslein, Tweepy: Twitter for Python! *URL: https://github.com/tweepy/tweepy* (2020).

84. E. Summers, *twarc: Archive tweets from the command line* (https://github.com/docnow/twarc).

85. M. W. Kearney, rtweet: Collecting and analyzing Twitter data. *Journal of Open Source Software*. **4**, 1829 (2019).

86. P. E. Shrout, J. L. Fleiss, Intraclass correlations: uses in assessing rater reliability. *Psychological bulletin*. **86**, 420–428 (1979).

87. Nitesh V. Chawla, K. W. Bowyer, L. O. Hall, SMOTE: Synthetic Minority Over-sampling Technique Nitesh. *Journal of Artificial Intelligence Research*. **2009**, 321–357 (2006).

88. N. Anand, D. Goyal, T. Kumar, in *Proceedings of International Conference on Recent Advancement on Computer and Communication* (2018), pp. 213–221.

89. H. Saif, M. Fernández, Y. He, H. Alani, in *LREC 2014, Ninth International Conference on Language Resources and Evaluation. Proceedings.* (2014), pp. 810–817.

90. M. F. Porter, An algorithm for suffix stripping. *Program*. **14**, 130–137 (1980).

91. G. A. Miller, WordNet: A Lexical Database for English. *Commun. ACM*. **38**, 39–41 (1995).

92. J. Li, X. Chen, E. Hovy, D. Jurafsky, in *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* (Association for Computational Linguistics, San Diego, California, 2016), pp. 681–691.

93. U. Michelucci, *Applied deep learning: A case-based approach to understanding deep neural networks* (Apress, New York, NY, 2018).

94. T. Mikolov, I. Sutskever, K. Chen, G. Corrado, J. Dean, in *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2* (Curran Associates Inc., USA, 2013), *NIPS'13*, pp. 3111–3119.

95. S. Hochreiter, J. Schmidhuber, Long Short-Term Memory. *Neural Computation*. **9**, 1735–1780 (1997).

96. P. Zhou, Z. Qi, S. Zheng, J. Xu, H. Bao, B. Xu, in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers* (The COLING 2016 Organizing Committee, Osaka, Japan, 2016), pp. 3485–3495.

97. K. Cho, B. van Merrienboer, C. Gulcehre, D. Bahdanau, F. Bougares, H. Schwenk, Y. Bengio, in *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing* (Association for Computational Linguistics, Doha, Qatar, 2014), pp. 1724–1734.

98. J. Chung, C. Gulcehre, K. Cho, Y. Bengio, in *NIPS 2014 Workshop on Deep Learning* (2014).

99. J. D. Kelleher, B. MacNamee, A. D'Arcy, *Fundamentals of machine learning for predictive data analytics: Algorithms, worked examples, and case studies* (MIT Press, Boston, Massachusetts, 2015).

100. S. Moreira, J. Filgueiras, B. Martins, F. Couto, M. J. Silva, in *Second Joint Conference on Lexical and Computational Semantics (*{SEM}), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation* (Atlanta, Georgia, 2013), pp. 490–494.

101. A. B. Warriner, V. Kuperman, M. Brysbaert, Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior Research Methods* (2013), doi:10.3758/s13428-012-0314-x.

102. A. Go, R. Bhayani, L. Huang, *Twitter sentiment classification using distant supervision* (2009), vol. 150.

103. L. F. Barrett, J. A. Russell, *The psychological construction of emotion* (The Guilford Press, New York, NY, 2014).

104. Y. R. Tausczik, J. W. Pennebaker, The Psychological Meaning of Words: LIWC and Computerized Text Analysis Methods. *Journal of Language and Social Psychology*. **29**, 24–54 (2009).

105. L.-P. Jing, H.-K. Huang, H.-B. Shi, in *Proceedings. International Conference on Machine Learning and Cybernetics* (2002), vol. 2, pp. 944–946 vol.2.

106. A. Liptak, Supreme Court revives transgender ban for military service. *The New York Times* (2019).

107. S. Deb, Accused of faking own assault, Jussie Smollett arrested on felony charge. *The New York Times* (2019).

108. P. Libbey, Sentiments on social media evolve with Jussie Smollett news. *The New York Times* (2019).

109. S. Mervosh, Viral video shows boys in 'Make America Great Again' hats surrounding Native Elder. *The New York Times* (2019), (available at https://www.nytimes.com/2019/01/19/us/covington-catholic-high-school-nathan-phillips.html).

110. H. Yan, C. Zdanowicz, E. Grinberg, Backlash erupts after United passenger gets yanked off overbooked flight. *CNN*.

111. J. Moffitt, *twitterdev/search-tweets-python* (@TwitterDev, 2021; https://github.com/twitterdev/search-tweets-python).

112. A. Tornes, Introducing Twitter premium APIs (2017), (available at https://blog.twitter.com/developer/en_us/topics/tools/2017/introducing-twitter-premium-apis.html).

113. G. M. Weiss, F. Provost, Learning when Training Data Are Costly: The Effect of Class Distribution on Tree Induction. *Journal of Artificial Intelligence Research*. **19**, 315–354 (2003).

114. S. Lei, H. Zhang, K. Wang, Z. Su, in *International Conference on Learning Representations* (2019).

115. R. Therrien, S. Doyle, in *Proc.SPIE* (2018), vol. 10581.

116. W. Pan, Akaike's information criterion in generalized estimating equations. *Biometrics* (2001), doi:10.1111/j.0006-341X.2001.00120.x.

117. W. J. Brady, J. A. Wills, J. T. Jost, J. A. Tucker, J. J. Van Bavel, Emotion shapes the diffusion of moralized content in social networks. *Proc Natl Acad Sci USA*. **114**, 7313–7318 (2017).

118. W. J. Brady, J. J. Van Bavel, Social identity shapes antecedents and functional outcomes of moral emotion expression in online networks (2021), , doi:10.31219/osf.io/dgt6u.

119. W. J. Brady, J. A. Wills, D. Burkart, J. T. Jost, J. J. Van Bavel, An ideological asymmetry in the diffusion of moralized content on social media among political leaders. *Journal of Experimental Psychology: General*. **148**, 1802–1813 (2019).

Acknowledgments

Acknowledgments: We thank members of the Yale Crockett Lab for valuable feedback throughout the project. We thank Kevin Hu and Mark Torres for assistance with fine-tuning of our outrage classifier. We thank Alan Gerber and Nicholas Christakis for providing helpful feedback on the paper. We thank the following students who contributed to annotating tweets: Mitchell Brown, Jonathan Burton, Vanessa Copeland, Vivian Fung, Aden Goolsbee, Sara Hollander, Berkeley Kijsriopas, Nikolette Lipsey, Sean Rice, India Robinson, Jordan Wylie, Lillian Yuan, Anna Zheng, and Michael Zhou.

Ethics statement: All research was conducted in accordance with the Yale University Institutional Review Board (IRB nos. 200026899 and 2000022385). For Studies 1 and 2, data collection was ruled "exempt" due to our use of public tweets to be aggregated into summary statistics when posted online. A public tweet is a message that the user consents to be publicly available rather than only to a collection of approved followers. The main potential risk of our research is that the users whose data we analyzed could be identified. To prevent this risk and ensure user privacy, data made available for other researchers is

in summary form only, without original text of tweets. This prevents anyone either accidentally or intentionally making public identifiable information of users if the data are reanalyzed. Furthermore, as stated in Table captions, all example tweets have been modified to prevent reverse text searches. In other words, people reading the paper cannot put the example text into google and find the user who posted the message. The broad benefit of this research is the production of knowledge that helps our society understand how the interaction of human psychology and platform design can shape moral and political behaviors the digital age. Over time, this knowledge can help inform scientific theory, policymakers and the general public.

Funding: This project was supported by the National Science Foundation, award #1808868, the Democracy Fund, award #R-201809-03031, and a Social Media and Democracy Research Grant from the Social Science Research Council.

Author contributions: W.J.B. and M.J.C. designed research; W.J.B., K.M., and T.N.D. performed research; W.J.B., K.M. and M.J.C. planned analyses; W.J.B analyzed data; W.J.B., K.M. and M.J.C. wrote the paper and all authors contributed to revisions.

Competing interests: Authors declare no competing interests

Data and materials availability: All data needed to evaluate the conclusions in the paper are present in the paper and/or the Supplementary Materials. All deidentified data, analysis scripts and preregistered data collection and analysis plans are available at https://osf.io/9he4n/

Figures and Tables

Fig. 1. Distributions of ideological extremity of user networks and levels of outrage expression. Panel A displays density plots of the ideological extremity of user networks for the Kavanaugh dataset (Study 1) and United dataset (Study 2). The x axis represents a continuous estimate of the mean ideological extremity of a user's network, greater values represent greater extremity. Panel B displays each user's median probability of expressing outrage in their tweets as a function of the ideological extremity of their network.

Fig. 2. Network-level ideological extremity moderates the effect of social feedback on outrage expressions. Each point displays the effect size estimate of previous social feedback predicting current outrage expression. Error bars were calculated based on standard errors of the estimate. The X axis represents quantile breaks from 20 to 80 percent. The blue color represents the Kavanaugh dataset users (Study 1), and the orange color represents the United dataset users (Study 2).

Fig. 3. Depiction of social media learning task (Studies 3 and 4). Participants first viewed what types of expressions were normative in their network by scrolling through 12 tweets. Next, they participated in a learning task where their goal was to maximize feedback.

Fig. 4. Reinforcement learning and norm learning shape outrage expressions in a simulated social media environment. The y axis represents the percentage of participants on each trial that selected outrage tweets to post. The x axis represents

the trial number. The red line represents participants in the outrage norm condition while the grey line represents participants in the neutral norm condition. Error bands represent the standard errors produced by fitting with a GAM function in R 3.6.1.

Table 1. Characteristics of all training datasets. DOC was first trained on 16,000 tweets collected during the Brett Kavanaugh confirmation hearings. We then tested generalizability and re-trained on the combination of Kavanaugh and all other topics (26,000 total tweets).

Table 2. Example outrage and non-outrage tweets as classified by the Digital Outrage Classifier. The table shows two example tweets from five political topics appearing in our training set that were classified as containing outrage vs. not containing outrage by DOC. To protect Twitter user privacy from reverse text searches, for figure display purposes only some words from each original tweet have been edited while maintaining salient features of the message.

Supplementary Materials

Supplementary Materials are attached in a separate document.