nature human behaviour

Article

https://doi.org/10.1038/s41562-023-01582-0

Overperception of moral outrage in online social networks inflates beliefs about intergroup hostility

Received: 16 September 2022

Accepted: 8 March 2023

Published online: 10 April 2023

Check for updates

William J. Brady[1,2] ✉, Killian L. McLoughlin[2,3,4], Mark P. Torres[5], Kara F. Luo[2], Maria Gendron[2] & M. J. Crockett[2,3,6] ✉

As individuals and political leaders increasingly interact in online social networks, it is important to understand the dynamics of emotion perception online. Here, we propose that social media users overperceive levels of moral outrage felt by individuals and groups, inflating beliefs about intergroup hostility. Using a Twitter field survey, we measured authors' moral outrage in real time and compared authors' reports to observers' judgements of the authors' moral outrage. We find that observers systematically overperceive moral outrage in authors, inferring more intense moral outrage experiences from messages than the authors of those messages actually reported. This effect was stronger in participants who spent more time on social media to learn about politics. Preregistered confirmatory behavioural experiments found that overperception of individuals' moral outrage causes overperception of collective moral outrage and inflates beliefs about hostile communication norms, group affective polarization and ideological extremity. Together, these results highlight how individual-level overperceptions of online moral outrage produce collective overperceptions that have the potential to warp our social knowledge of moral and political attitudes.

Functional democracies require citizens to acquire accurate social knowledge about collective moral attitudes[1,2]. For instance, resolving how a society can balance the right to free speech against the harms caused by hate speech requires a shared understanding of moral attitudes regarding freedom of expression and harmful speech. If individuals overperceive how morally wrong others view an issue, this inaccurate social knowledge could hinder progress on finding the common ground required for effective cooperation since individuals may not understand which issues an opposing group actually cares the most about[3]. In the digital age, social interactions increasingly occur in the context of online social networks and political leaders frequently use them as a tool for communication. Thus, it is important to understand how online social network platforms can shape social knowledge of morality and politics.

Recent work has argued that social media platforms—as they are currently designed—can distort social knowledge of morality and politics[1,4–6]. While empirical work in this area has mainly focused on the role of platforms in spreading disinformation, here we examine how they may exacerbate a basic psychological bias in social perception: overperception of negatively valenced emotions. We focus on moral outrage—a mixture of anger and disgust triggered by a perceived moral norm violation[7,8]—in particular because of its key role in signalling to

[1]Kellogg School of Management, Northwestern University, Evanston, IL, USA. [2]Department of Psychology, Yale University, New Haven, CT, USA. [3]Department of Psychology, Princeton University, Princeton, NJ, USA. [4]School of Public and International Affairs, Princeton University, Princeton, NJ, USA. [5]Department of Statistics and Data Science, Yale University, New Haven, CT, USA. [6]University Center for Human Values, Princeton University, Princeton, NJ, USA. ✉e-mail: william.brady@kellogg.northwestern.edu; mj.crockett@princeton.edu

others that a morally relevant event or action has occurred[7,9] and its role in motivating collective action and political behaviour[10,11] (henceforth we use 'outrage' interchangeably with 'moral outrage'). We propose that communication on social media is conducive to overperception of individuals' moral outrage, which contributes to overperception of collective moral outrage.

We consider that overperception of moral outrage on social media can arise from a complex interplay of factors that affect both observers' perceptual processes as well as authors' motivations to express outrage. On the observer side, several features of social media platforms have the potential to create overperception of individuals' moral outrage. First, observers perceive authors' emotional expressions through limited channels (text/images) that lack the richer informational cues that typically accompany real-life emotion expressions and are important for accurate emotion perception[12]. Past work demonstrates that perceiving emotions on the basis of a restricted set of linguistic or graphic cues in computer-mediated environments specifically leads to overperception of negativity[13–15]. This occurs because limited cues and lack of real-time social feedback create ambiguous communication intent and under such conditions false alarms are less costly than misses[14,15]. Furthermore, when observers believe that social media environments have high levels of outrage—either through popular media narratives or actually seeing a lot of outrage in their feeds because such content is promoted by content algorithms—they may form rational priors that increase their likelihood of perceiving outrage expressions on the platforms[9,16,17]. Together, these factors suggest that social media users are likely to overperceive moral outrage, especially those users who have spent more time in online networks where outrage is common and thus have had more time to form prior beliefs about the prevalence of outrage.

For authors, social and group identity motivations that are particularly salient on social media platforms, combined with social learning processes, might encourage people to express moral outrage more frequently or more intensely than they actually feel[9,16,18]. For example, expressing outrage serves as a signal of group affiliation and trustworthiness that can enhance a person's reputation[9,19,20]. Social media environments amplify these reputational rewards because they provide a much larger audience for outrage expressions than is typical for most people offline[21]. Moreover, outrage expressions tend to be highly rewarded by the social feedback delivery system inherent to platforms[16,22], which causes users to learn to express more outrage over time. Thus, some authors on social media may be motivated to express outrage in ways that are not tethered to their actual experiences of outrage. Alongside the platform features that affect the accuracy of observer perceptions, these incentives for authors create a recipe for overperception of outrage online.

Overperception of individuals' outrage may amplify perceptions of collective outrage at the group level. Through the lens of classical models of social network learning[23], overperceiving several individual outrage expressions may amplify perceptions of collective outrage because people take the average of the individual expressions to gauge how the group feels collectively. In other words, biased individual judgements will lead to a biased average judgement. In addition, social media environments display viral emotion expressions alongside one another in a newsfeed, so that people view multiple emotional expressions in their network at the same time. This feature is consequential because perceiving emotions of many group members at once makes people overperceive the extremity of group emotions (the 'crowd-emotion-amplification-effect'[24]). More generally, social media newsfeed algorithms tend to push evocative content because it draws more engagement and this creates a situation in which a minority of politically extreme social media users create the majority of political content that people see[25–29]. Consequently, people are presented with a biased sample of social content that they use to form impressions of their networks' collective feelings. Together, these findings suggest that by being conducive to overperception of individuals' outrage, social media can further amplify perceptions of collective outrage.

Documenting the overperception of individual and collective outrage (if it occurs) can shed light on a key process by which social media distorts our knowledge of morality and politics with consequences for intergroup relations. When we overperceive how outraged others in our network are, we may also increase our belief that it is socially appropriate to express outrage in the network (norms of outrage expression), that our network dislikes the political outgroup (affective polarization) and that our network is politically extreme (ideological extremity). These beliefs can be problematic for intergroup relations because people often conform to social norms even when they are overperceived[30,31]. For instance, recent work suggests that when people overperceive the extremity of a group's moral attitudes, it leads them to adopt more extreme attitudes themselves[32–35]. Understanding the affective building blocks of intergroup misperceptions can advance theories of intergroup relations for the digital age.

Results

To test for overperception of outrage and its consequences for intergroup outcomes, we developed a methodology that allowed us to measure the outrage felt by social media users when they posted a message and then compare it to observer judgements of the authors' outrage. Across three field studies, we find evidence for systematic overperception of individuals' outrage on Twitter; namely, that observers perceive more outrage than is reported by authors. In a preregistered confirmatory experiment using a simulated Twitter newsfeed, we manipulated overperception of individuals' outrage and found that when participants view a newsfeed containing outrage expressions that tend to be overperceived, it causes perceptions of collective outrage of the social network to increase. A second preregistered experiment found that viewing overperceived outrage messages also amplified participants' beliefs about (1) norms of outrage expression, (2) affective polarization and (3) ideological extremity present in the network. Together, these results shed light on the social dynamics of moral outrage in online networks that can exacerbate polarization as our social lives become more digital than ever before.

Twitter field studies examining overperception of outrage

To test for overperception of outrage on social media, we conducted Twitter field studies consisting of two phases: an author phase and an observer phase. In the author phase, we used machine learning to identify Twitter users (authors) who expressed high or low levels of outrage during Twitter conversations about contentious American political topics[16] (Table 1). Shortly after they posted the tweets (within 15 min), we invited authors to report how outraged and happy they felt when they composed their tweets. In the observer phase, we recruited an independent group of politically partisan American social media users to view the tweets of the authors from the author phase (observers) and judge how outraged and happy they believed the authors were when they posted the tweets. See Methods for more details and Fig. 1 for an overview of the field study method. We chose to study Twitter because it is known to host large-scale episodes of public moral outrage[36] and many important political and public figures use it to communicate with their audiences.

The first field study ($N = 133$ authors, $N = 110$ observers) was conducted in July and August 2020 and examined tweets discussing the William Barr Congressional hearing as well as President Trump's favourability. The second preregistered, confirmatory field study ($N = 200$ authors, $N = 190$ observers) was conducted in October 2020 and examined tweets discussing the Amy Coney Barrett Supreme Court confirmation and the 2020 US Presidential election (Methods).

For our main analysis, we tested whether observers tended to overperceive authors' moral outrage. We conducted a multilevel model that included random effects of target (tweet) and observer. In the

https://doi.org/10.1038/s41562-023-01582-0

Table 1 | Data collection details

Study	Dates active	Political topic	Keywords used	Topic description
Study 1	28 July to 3 August 2020	William Barr Congressional Hearings, President Trump favourability	'barr', 'barrhearings', 'removebarrnow', 'nobodylikestrump'	In a highly politicized congressional hearing, Attorney General William Bar is cross-examined by Democratic Senators regarding allegations of unprecedented legal interventions on behalf of President Donald Trump. Discussions of President Trump's favourability and behaviour were debated by partisans across the political spectrum with the context of the upcoming 2020 election in focus.
Study 1	12 to 18 August 2020	William Barr Congressional Hearings, President Trump favourability	'barr', 'barrhearings', 'removebarrnow', 'nobodylikestrump'	(as above)
Study 2	22 to 30 October 2020	Senate confirmation hearings for the nomination of Amy Coney Barrett to the US Supreme Court, US Senator behaviours	'coney barett', 'confirmation hearings', 'supreme court', 'court nomination', 'senate vote', 'feinstein', 'lindsey graham'	In a highly contentious confirmation hearing, Senate Democrats and Republicans questioned President Trump-appointed Supreme Court Nominee Amy Coney Barrett. The hearings would determine if she was fit to be a Supreme Court Judge. Her confirmation would change the balance of the Supreme Court to be much more conservative, which made it important for Republicans and threatening for Democrats.

The table lists the dates of data collection, political topics targeted and keywords provided to the stream API to collect tweets.

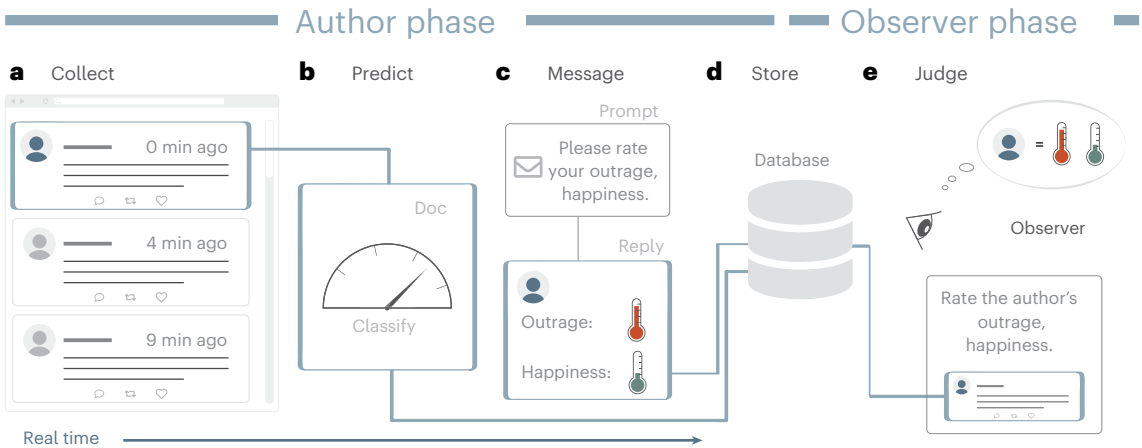

Fig. 1 | Overview of Twitter field study method. a–**e**, The study was organized in stages indicated as collect (**a**), predict (**b**), message (**c**), store (**d**) and judge (**e**). In the author phase, a DOC detected outrage expression in real time as people with public profiles tweeted about contentious topics in American politics (**a**,**b**). Next, users with open direct messaging were messaged from the Yale Social Media Research account asking if they would volunteer to self-report the outrage and happiness they felt when tweeting the message (**c**). In the observer phase, messages (with author information removed) were shown to a new group of Republican and Democrats participants (observers). Observers judged how outraged and happy each message author was using the same scale that message authors used to report their outrage and happiness in the author phase (**d**,**e**).

model, source (author versus observer) was entered as a dummy-coded predictor variable and tweet id and observer id and were entered as the clustering variables. This analysis revealed consistent evidence for overperception of outrage: observers reported perceiving higher levels of outrage than reported by message authors in study 1 (slope coefficient (b) = 0.59, P = 0.011, 95% confidence interval (CI) = (0.14, 1.04)) and study 2 (b = 0.58, P = 0.001, 95% CI = (0.25, 0.92)) (Fig. 2). While we observed consistent overperception of outrage, we did not observe overperception of happiness in study 1 (b = −0.13, P = 0.538, 95% CI = (−0.54, 0.28)) nor study 2 (b = −0.17, P = 0.295, 95% CI = (−0.49, 0.14)). These results were robust to various analytic strategies (Supplementary Section 1.2). We also verified that observers' ratings were not merely 'noise': they were significantly correlated with authors' reports but they tended to overestimate the amount of outrage for any given tweet (Supplementary Section 1.2). For exploratory analyses that disentangle author versus observer effects in the overperception finding and that examine partisanship differences see Supplementary Sections 1.3, 1.6 and 1.7.

As an exploratory analysis, we combined data from studies 1 and 2 to test whether observers' tendency to overperceive outrage was associated with their amount of daily social media use to learn about politics (political social media use). We reasoned that observers higher in political social media use would be exposed to more outrage in the context of politics (due to both user behaviours and upranking by content algorithms) and thus may have stronger prior beliefs about the normativity of outrage expression and be more likely to overperceive outrage than observers lower in political social media use. Consistent with this reasoning, we found a significant positive correlation between overperception of outrage and political social media use (r(222) = 0.19, P = 0.004, 95% CI = (0.06, 0.31)) (Fig. 3; Methods). Multiple regression analyses revealed that political social media use was a significant predictor of overperception of outrage when statistically adjusting for observers' ideological extremity, partisan identity strength and tendency to overperceive happiness (Supplementary Section 1.3). These findings suggest that the relationship between observers' overperception of outrage and their political social media use is not simply explained by the fact that frequent social media users are often more politically extreme or more strongly identified with their political group[6], although further studies are required to fully rule out this possibility.

https://doi.org/10.1038/s41562-023-01582-0

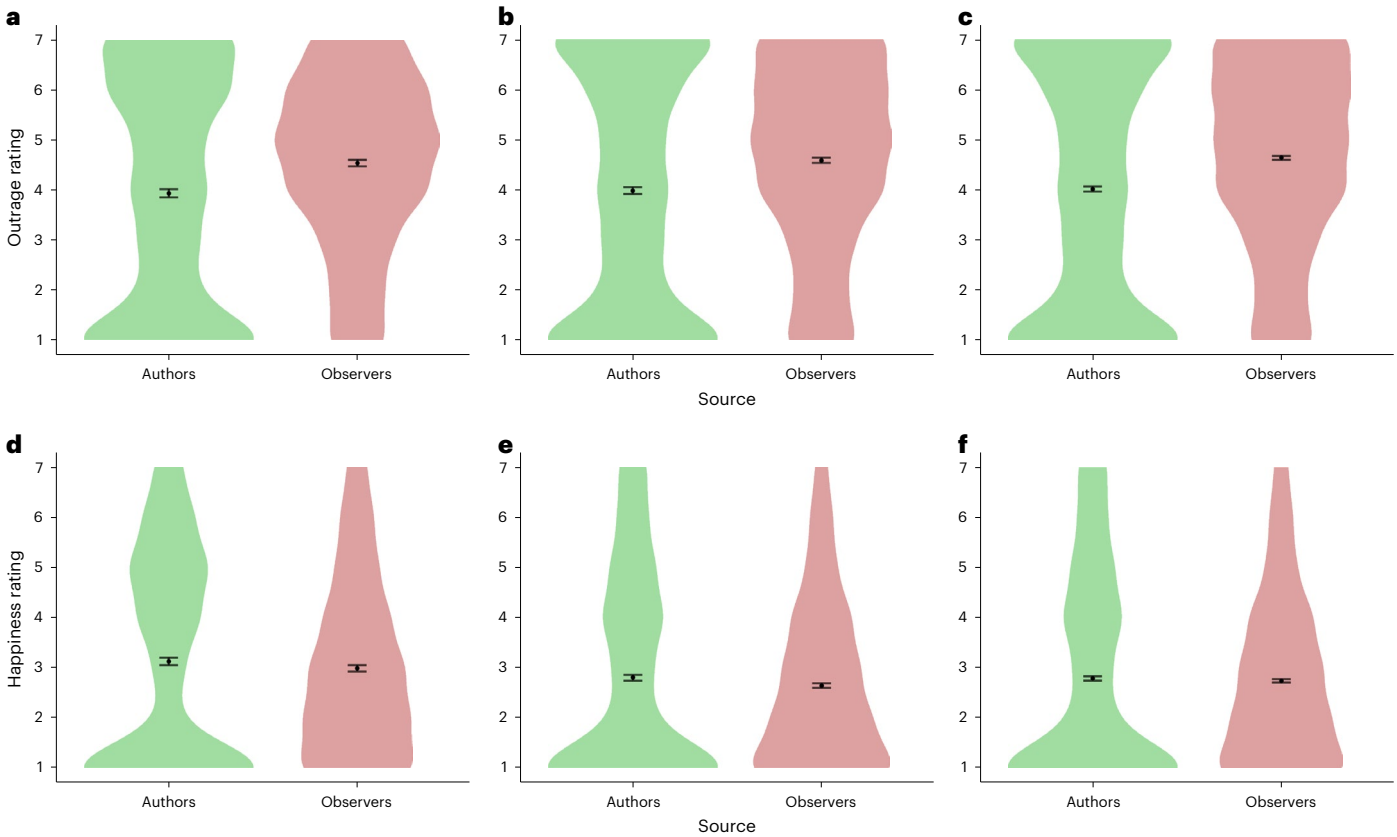

Fig. 2 | Evidence of overperception of individuals' moral outrage on Twitter. **a**–**c**, Outrage ratings self-reported by message authors (green) and the mean of ratings as judged by an independent group of observers (red) for studies 1–3 (**a**,**b**,**c**, respectively). In all studies, the amount of outrage perceived by observers was significantly greater than the outrage reported by message authors, demonstrating overperception. **d**–**f**, Happiness ratings self-reported by message authors (green) and the mean of ratings as judged by an independent group of observers (red) for studies 1–3 (**d**–**f**, respectively). In all studies, the amount of happiness perceived by observers was not significantly different than the happiness reported by message authors, demonstrating a selective case of overperception for moral outrage. Study 1, *N* authors = 133 and *N* observers = 110; study 2, *N* authors = 200 and *N* observers = 190; study 3, *N* authors = 200 and *N* observers = 350. Error bars, ±1 s.e.m.

Next, we conducted a third preregistered study to test whether the overperception of individuals' outrage would replicate for a third time and also to provide a confirmatory, high-powered replication of the relationship between overperception of outrage and political social media use. In study 3, we carried out the observer phase only and recruited a larger number of participants ($N = 350$) to make judgements about the author-phase tweets from study 2. Study 3 replicated the main overperception finding, with observers perceiving more outrage than was reported by authors, $b = 0.62$, $P < 0.001$, 95% CI = (0.27, 0.93). Once again, overperception occurred only for outrage and not for happiness judgements, $b = -0.06$, $P = 0.656$, 95% CI = (−0.50, 0.11).

Study 3 also replicated the association between overperception of outrage and political social media use, $r(248) = 0.20$, $P = 0.001$, 95% CI = (0.08, 0.32) (Fig. 3). Multiple regression analyses revealed that political social media use was a significant predictor of overperception of outrage when statistically adjusting for observers' ideological extremity and partisan identity strength, as well as their tendency to overperceive happiness (Supplementary Section 1.12).

Across studies 1–3, we found robust evidence that, at the individual level, observers perceived more outrage in messages than the authors of those messages reported actually feeling. This finding provides evidence that observers from across the political spectrum overperceive the level of outrage expressed in messages posted by politically active social media users. The overperception of individuals' emotions occurred for moral outrage but not for happiness. Furthermore, we found that observers' overperception of outrage is positively associated with their political social media use, suggesting that people who spend more time using social media to learn about politics are more likely to overperceive outrage.

Overperception of individual and collective outrage

Study 4 tested whether the overperception of individuals' outrage that we discovered in studies 1–3 amplifies perceptions of collective moral outrage. In other words, when people view a newsfeed containing multiple messages whose outrage expressions tend to be overperceived, does it cause observers to overperceive the total amount of outrage in the network as a whole?

For these experiments, we leveraged the database of author tweets that we gathered in studies 1 and 2 to create simulated Twitter newsfeeds. For each tweet in our database, we had access to the author's self-reported outrage, as well as the mean of at least ten observers' judgements of outrage (Fig. 4a). These pairs of author self-reports and observer mean judgements enabled us to construct two sets of tweets: 'high-overperception' tweets where the observers reported notably more outrage than the authors reported and 'low-overperception' tweets, where the observers' judged outrage and authors' self-report were within 1.5 scale points on average (Methods). From these tweets, we constructed two simulated Twitter newsfeeds: a high-overperception newsfeed composed of highly overperceived tweets and a low-overperception newsfeed composed of less overperceived tweets. Crucially, these feeds were matched in the level of outrage self-reported by tweet authors (Fig. 4b). This method allowed

https://doi.org/10.1038/s41562-023-01582-0

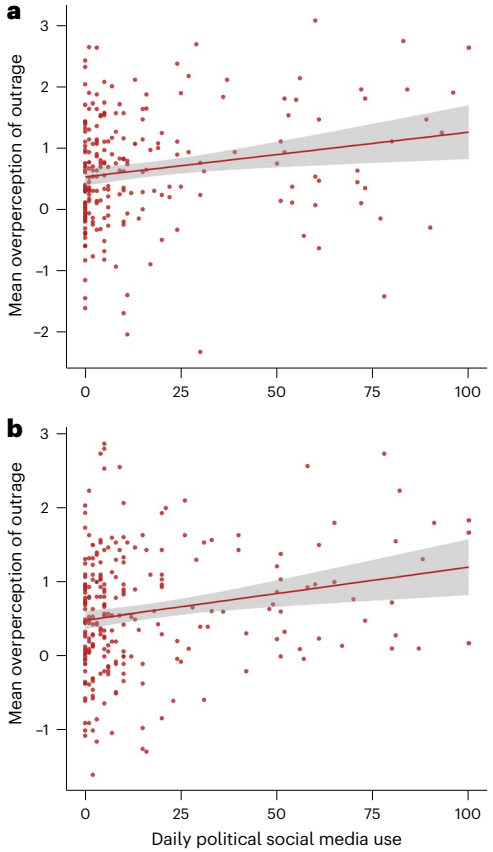

Fig. 3 | Observers' overperception of outrage is positively associated with their daily political social media use. a,**b**, Overperception of outrage was defined at the participant-level, where the *y* axis represents the mean levels of outrage judged by an observer minus the mean outrage reported by authors (for all messages the observer judged). Daily political Twitter use was measured on the basis of a slider scale from 0 to 100, where users who did not use social media daily to learn about politics were instructed to select '0'. **a**, Exploratory results from studies 1 and 2, two-sided Pearson correlation $r(222) = 0.19$, $P = 0.004$, $N = 224$. **b**, Confirmatory analysis from study 3, two-sided Pearson correlation $r(248) = 0.20$, $P = 0.001$, $N = 250$. Error bands represent 95% CIs on linear model predictions.

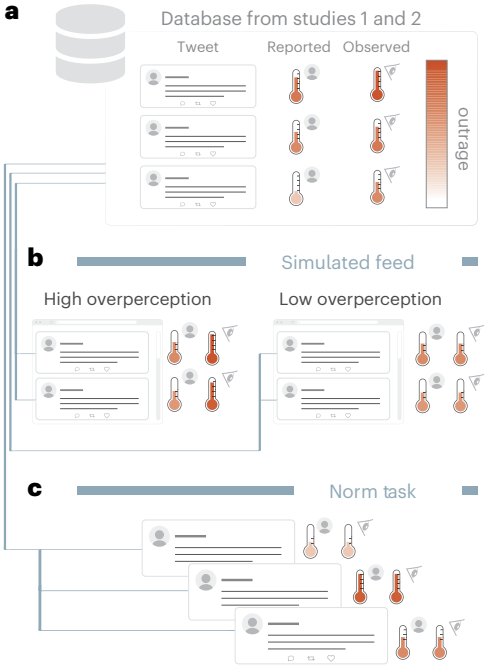

Fig. 4 | Depiction of newsfeed manipulation in a mock social media environment. a–**c**, In studies 4 and 5, participants were randomly assigned to one of two conditions: a high-overperception newsfeed or a low-overperception newsfeed. **a**, In each condition, tweets were pulled from studies 1 and 2. **b**, In the high-overperception condition, tweets that were highly overperceived (observer judgements were much greater than author reports) were displayed in the newsfeed; in the low-overperception condition, tweets that were more accurately perceived were displayed in the newsfeed (observer judgements were within 1 scale point to author reports on average). **c**, In the 'norm task', participants viewed new tweets and judged how socially appropriate each tweet would be to post to their social network.

us to test our hypothesis that when people are exposed to a group of tweets whose outrage tends to be more overperceived, they will perceive the collective moral outrage of the social network to be greater.

For the preregistered experiment, we randomly assigned participants ($N = 600$) to view either the high-overperception newsfeed or the low-overperception newsfeed in a between-subjects experiment. Participants were then asked to make a judgement of collective outrage (Methods). Our main hypothesis was that participants in the high-overperception newsfeed condition would judge the collective outrage of their social network as significantly greater than the low-overperception newsfeed condition, even though the authors' self-reported outrage was held constant in both conditions.

As expected, participants in the high-overperception newsfeed condition judged the collective outrage of their social network to be significantly greater (mean (M) = 5.82) than participants in low-overperception newsfeed condition ($M = 3.53$), $t(479.50) = 21.56$, $P < 0.001$, Cohen's $d = 1.90$, 95% CI = (2.09, 2.50) (Fig. 5). This finding supports our prediction that overperception of individuals' outrage directly amplifies perceptions of collective outrage.

In follow-up analyses, we examined the social learning process by which participants in the overperception newsfeed condition increased their perceptions of collective outrage. A 'simple average' prediction derived from classical models of social network learning[23] would be that participants' judgements of collective outrage should equal the mean of outrage perceived in each newsfeed message when viewed individually (a message's 'individually perceived outrage' value). Alternatively, a 'weighted average' prediction[23] would be that certain messages in newsfeeds factor into their collective judgement of outrage more than others and collective judgements will not perfectly equal the mean of individually perceived outrage values. We used the observer judgements from studies 1 and 2 to calculate the mean of individually perceived outrage values (Methods).

Supporting the weighted average prediction, we found that in the high-overperception condition, participants' perceptions of collective outrage ($M = 5.82$) were significantly greater than the mean of outrage perceived in each newsfeed message when viewed individually ($M = 5.30$), $t(264) = 8.06$, $P < 0.001$, $d = 0.49$, 95% CI = (5.69, 5.95). This effect was not found in the low-overperception condition, $t(252) = 1.38$, $P = 0.169$, $d = 0.09$, 95% CI = (3.36, 3.69). Since the collective outrage judgements were greater than the mean of individually perceived outrage values, our results suggest that participants in the high-overperception condition were weighting the most intense outrage messages more than less intense outrage messages when making their collective outrage judgements.

Consequences of overperceiving collective outrage

For study 5, we conducted a preregistered experiment ($N = 1,200$) to examine consequences of overperception of collective outrage. Using the same manipulation as in study 4, participants were randomly assigned to either the high-overperception newsfeed or the

https://doi.org/10.1038/s41562-023-01582-0

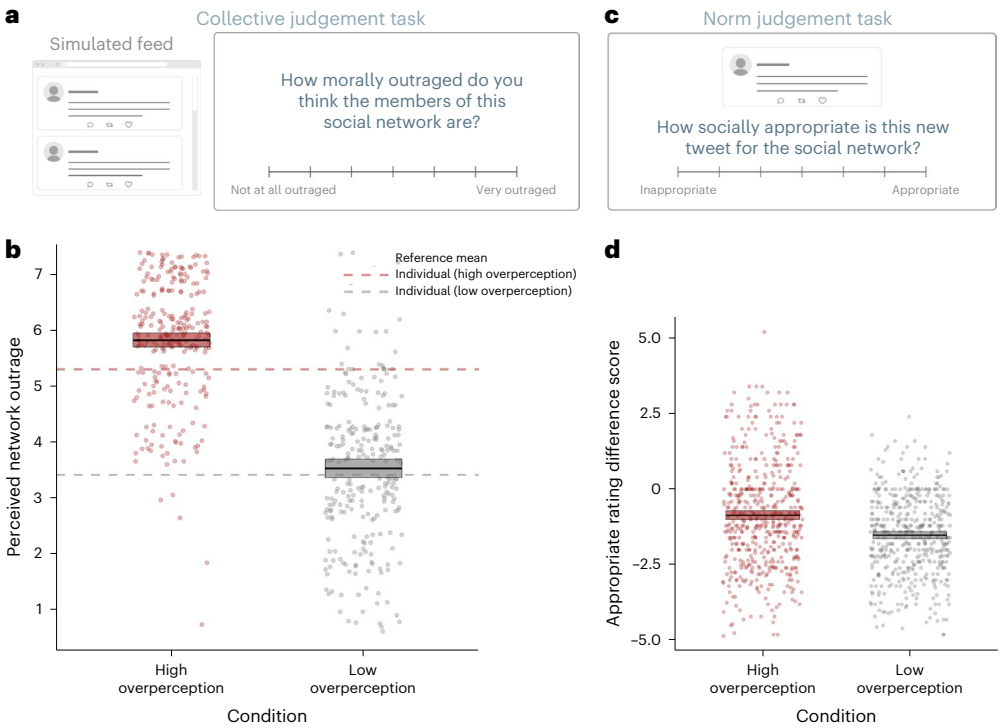

Fig. 5 | Overperception of outrage in newsfeeds amplifies perceptions of collective outrage and beliefs about outrage norms. a, In Study 4, participants viewed a simulated newsfeed with the overperception manipulation depicted in Fig. 4. and then judged how collectively outraged members of the social network were ('collective judgment task'). **b**, Study 4 found that participants in the high-overperception condition ($n = 265$) judged their social network to be more collectively outraged than participants in the low-overperception condition ($n = 258$). In the high-overperception condition, participants' collective outrage judgement was even greater than the mean of outrage judged in each individual message in studies 1 and 2 (red dotted line; **b**). In the low-overperception condition, participants' collective outrage judgement was also greater than the mean overperception factor of each individual message in their newsfeed (grey dotted line; **b**), although the effect was much smaller. **c**, In Study 5, participants again viewed a simulated newsfeed with the overperception manipulation, but viewed new tweets and were asked how socially appropriate the tweet would be for the social network ('norm judgement task'). **d**, Study 5 found that viewing the high-overperception newsfeed ($n = 523$) led to greater endorsement of new outrage messages as socially appropriate compared to the low-overperception newsfeed ($n = 490$). Boxplots, means ± 1 s.e.m.

low-overperception newsfeed (Fig. 4 and study 4). In study 5, however, after viewing the newsfeed, participants were asked about their perceptions of outrage norms, affective polarization and ideological extremity of the social network (all findings reported below assume a Bonferroni-corrected $d = 0.017$ to account for multiple comparisons).

To examine perceptions of outrage norms, after viewing the simulated newsfeed, participants completed a norm judgement task where they were exposed to ten new political tweets that expressed political opinions with either outrage or neutral language (see Methods for details on tweet selection for this task). Participants were then asked to judge how socially appropriate each tweet would be to post to the social network they had viewed in the newsfeed. We found that participants exposed to the high-overperception newsfeed judged tweets that expressed outrage as more socially appropriate (relative to more neutral tweets) than participants who were exposed to the low-overperception newsfeed, $t(964.58) = -6.89$, $P < 0.001$, $d = 0.43$, 95% CI = (0.47, 0.85) (Fig. 5; Methods). Thus, viewing a newsfeed that produced overperception of collective outrage amplified people's perceptions of how normative it is to express outrage in the social network.

Next, we investigated whether overperception of outrage inflates beliefs about affective polarization. To measure affective polarization, we asked participants to judge how their social network felt about the political ingroup and outgroup using a feeling thermometer measure (Methods). As revealed by a significant interaction between newsfeed condition and group, $F(1, 1{,}010) = 369.58$, $P < 0.001$, $\eta_p^2 = 0.27$, participants assigned to view the high-overperception newsfeed judged their network to like the political ingroup more ($M_{\text{ingroup}} = 75.07$) compared to those who viewed the low-overperception newsfeed ($M_{\text{ingroup}} = 56.12$), $P < 0.001$. Participants assigned to view the high-overperception newsfeed also judged that their network disliked the political outgroup more ($M_{\text{outgroup}} = 15.64$) than those who viewed the low-overperception newsfeed ($M_{\text{outgroup}} = 48.23$), $P < 0.001$. Post hoc comparisons revealed that the effect of newsfeed condition on outgroup dislike ($M_{\text{diff}} = 32.59$) was nearly twice as large as the effect on ingroup liking ($M_{\text{diff}} = 18.95$). These results suggest that overperceiving outrage in the newsfeed notably increases the belief that the social network is affectively polarized.

Lastly, we asked participants in both newsfeed conditions to judge how ideologically extreme people in the social network were. We found that participants who viewed the high-overperception newsfeed judged their network to be more ideologically extreme ($M = 1.89$) than did participants who viewed the low-overperception newsfeed ($M = 1.34$, $t(1003.60) = -11.39$, $P < 0.001$, $d = 0.72$, 95% CI = (0.46, 0.64)). Taken together, the findings of study 5 suggest that overperception of outrage in newsfeeds directly amplifies perceptions of norms of outrage expression, affective polarization and ideological extremity in social networks.

Discussion

Across exploratory and confirmatory Twitter field studies, we found social media users tend to overperceive the moral outrage authors express in their messages, while we found no evidence of overperception when they perceived happiness. The discrepancy between the moral outrage ratings of authors and observers was greatest for users who had the highest daily political social media use. In preregistered

follow-up experiments, we found that overperception of individuals' moral outrage amplifies perceptions of collective moral outrage and beliefs about norms of moral outrage expression, affective polarization and ideological extremity.

These results provide evidence that perceiving emotions in a constrained social media environment, combined with expectations formed from using social media often, is associated with overperception of individuals' moral outrage. Future work should investigate further which specific features of social media explain the most variance in producing overperception of various emotions. For instance, a key feature of social media is the limited amount of communication cues present in language and images that inform emotion perception in computer-mediated communication[9,12,37]. Another feature of social media is that emotion perception occurs more asynchronously compared to offline contexts[14], which removes the opportunity for an expressor to provide live feedback that update initial perceptions. Future work should also investigate how network-level features of social media can produce overperception of emotions, such as the directed nature of the social networks formed, high levels of homophily and the tendency for people to have fewer friends than their friends have (the 'friendship paradox')[38,39]. Further research is also required to compare the extent to which these features of social media create the tendency for other emotions besides moral outrage to be overperceived. For instance, because negative emotions spread disproportionately on social media platforms[22,40], their frequent presence may create expectations of negativity that influence perception.

We also found that overperception of individuals' outrage amplifies perceptions of collective outrage. Furthermore, when making collective outrage judgements, people appear to weight more intense outrage expressions more than less intense outrage expressions to form their collective judgement. These results suggest that biases impacting emotion perception of individuals translate into even larger biases when judgements are made about a group. This finding sheds light on the affective building blocks of large, collective biases such as in cases of pluralistic ignorance[30,31] where attitudes held by a minority of group members are believed to be the majority. One process that may explain this finding is the 'crowd-emotion-amplification effect' or the idea that group settings amplify emotion perception biases because people's attention is drawn to extreme emotion expressions in group settings[24]. This may be particularly exacerbated on social media as emotion expressions are often displayed alongside one another in a newsfeed and outrage expressions tend to draw engagement and influence users' understanding of emotion norms in the social network[9,16]. However, future research is required to disentangle the effects of 'crowd-emotion-amplification' versus the impact of making individual versus group judgements more generally.

Finally, we found that overperception of outrage has important consequences for network-level social perceptions: inducing overperception of outrage amplified perceptions of norms of outrage expression, affective polarization and ideological extremity within the social network. These findings shed light on the key affective processes in social perception that may underlie actual polarization in social networks. If people perceive their social network to be more outraged at the outgroup than they really are, then preference falsification[41] may occur where people express more outrage than they actually feel to conform to the perceived majority. On the other hand, outrage might go unchecked in a 'spiral of silence'[42,43], where more moderate individuals are hesitant to express opposition to the outrage in their network because they perceive it to be the majority emotion. Each of these social processes can amplify political polarization since outrage toward outgroups will become amplified beyond the true base-rates of outrage felt by each individual in the social network. Indeed, recent work on 'false polarization' suggests that people become more polarized when they mistakenly believe that political groups have more extreme attitudes than they actually do[33–35]. Overall, our work suggests that examining the conditions under which moral emotions are prone to overperception can help identify when false polarization is most likely to occur and when more extreme voices are asymmetrically represented in a social network.

Our results also raise the possibility that correcting people's ingroup perceptions by providing them with more accurate social information about the ingroup's underlying emotions could help combat the tendency to overperceive—or even conform to—perceived norms of outrage expression. For example, previous work has shown that providing people with more accurate information about an outgroup's beliefs can help mitigate inaccurate metaperceptions of the outgroup[44]. However, in the case of social media, social perception takes place in a technologically mediated environment and the social information people witness may be skewed based on algorithmic behaviour that promotes more extreme voices[6,9]. If a user continues to view biased social information every time they log on, they might ignore factual, base-rate information given to them previously. On the other hand, an effective route to correcting overperception of outrage on social media may be to design an educational intervention to make them aware that platform algorithms can skew what appear to be representative emotions or attitudes in their network. Recent work suggests that people dislike the idea that algorithms can alter social information in their social media experience, suggesting that people may be motivated to learn more about how algorithms can skew social information[45,46].

This work has several limitations. First, our results in studies 1–3 are based on the selection of specific Twitter users whom we were able to contact and who responded to our messages. This raises concerns that this particular group of users were disproportionately likely to overexpress their outrage, which would make our results less generalizable to other Twitter users. However, in a series of robustness tests (Supplementary Section 1.5), we found that the authors in our field studies expressed slightly less outrage overall in their tweet history relative to comparison groups, suggesting that the present findings are actually a conservative test of our hypotheses. In addition, these users did not systematically differ in their political extremity, Twitter use or follower count compared to users who did not respond to our direct message (DM) and users who tweeted about a different political topic. These tests suggest that our findings cannot be explained away by a selection bias and may generalize beyond the political topics studied here (Supplementary Section 1.5). Future work is required to test the extent to which overperception of moral outrage extends to message authors who are less politically active but it is noteworthy that on social media more extreme political content that spreads widely is often produced by politically active users who are a minority on the platform as a whole[25]. Even if more extreme users are not necessarily representative of all users, they may have an outsized influence in terms of message diffusion[26].

Another limitation is that, in the observer phase of studies 1–3, participants made judgements of the outrage in tweets outside of the full social media context. While this allowed us to hone in on the specific role of language in emotion perception, on social media there are other cues that may influence perception including the author's social identity, the history perceivers have with the author and context effects including the amount of outrage surrounding any given message. All these factors make perception of emotions on social media platforms more complicated and these factors may all interact to influence perception alongside the expression of emotion embedded within message text itself. Furthermore, our finding that overperception of outrage is positively correlated with political social media use suggests that a perceiver's history on the platform is also an important factor to consider when understanding the factors contributing to overperception of outrage. Further research is required to determine whether our results hold for other platforms that may have different communication norms, users and algorithm behaviour or even other

media channels (for example, television)[12]. Further research is also required to determine the extent to which our results will apply to less political or moral topics that may come with different communication incentives and social rewards[9,21]

We were also unable to fully discern whether overperception of individuals' outrage in studies 1–3 was driven more by biases present in the observer that make them overperceive otherwise accurate emotion information or biases present in the author that lead them to express exaggerated signals of their emotional feelings. However, we found that observers' political social media use predicted overperception, whereas several author characteristics such as their social reinforcement history did not, suggesting that observer priors are a key component of overperception (Supplementary Section 1.3). Above, we articulated several reasons why both sorts of biases can be amplified on social media due to an interaction of group psychology and constraints of the social media environment (see also ref. 9). While assumptions about author versus observer effects in biased communications tends to vary by discipline[30], our best understanding of emotion perception bias on social media is likely to come from models that fully integrate both the author and observer as sources of bias (for example, ref. 47).

Finally, in studies 4 and 5 we were able to examine how high overperception of individuals' outrage affected judgements of collective outrage relative to low overperception of individuals' outrage. However, these studies were not well-equipped to examine the absolute accuracy of group-level judgements given that we had truth criteria for individuals in the group (self-reports of outrage) but no obvious truth criteria for the group as a whole. Future studies should examine experimental designs that can better tease apart individual versus group-level accuracy, such as in recent work on group metajudgements[32]. Furthermore, it would be interesting in future work to tease apart the role of collective judgements of outrage versus individual judgements of outrage (for example, when people are exposed to single Twitter messages) in predicting perceptions of norms and polarization.

Across Twitter field studies and behavioural experiments, we found that people tend to overperceive moral outrage in political social media messages and that this overperception amplifies perceptions of collective outrage and beliefs about outrage norms, polarization and ideological extremity. Our results provide a starting point for understanding the psychological processes that distort social knowledge of moral and political information on social media as the platforms are used more than ever before to learn about morality and politics. Our findings suggest that one of the key challenges of social media platforms is promoting accurate social perceptions by preventing the most extreme moral and political content from being over-represented in people's social media experience.

Methods

Twitter field studies (studies 1–3)

All field studies described below were approved by Yale University's IRB, approval no. 2000026899. Author-phase participants were informed that if they responded to our DM, their responses would be used for our research and would remain anonymous. Observer-phase participants consented to research by agreeing to an online informed consent form. Our field studies were conducted on the basis of a research pipeline we built that is described in Fig. 2. The author phase consisted of three stages. First, we searched Twitter's application programming interface (API) for public Twitter messages (tweets) about contentious topics in American politics. Second, we classified the tweets found in the search for whether they contained moral outrage expression. Third, we direct messaged users who consented to open DMs and asked them to report on the emotions they experienced when they posted their tweet.

To search Twitter for political topics and outrage tweets, we connected to Twitter's standard API and leveraged the streaming endpoint via Python v.3.8. We first streamed tweets containing keywords to find tweets about contentious political topics. We targeted the time of data collection around contentious events that unfolded in American politics (Table 1). Study 1 served as a proof of concept for our research pipeline and main hypothesis and we had to collect data in two waves due to an error in data collection during the first wave. Study 2 served as a confirmatory test that we preregistered with predetermined political topics and dates for data collection (https://osf.io/ud5bc).

As tweets were being collected from the API, we then classified each of them with the Digital Outrage Classifier (DOC)[16] to determine whether they contained moral outrage expression. DOC was trained on the basis of training datasets labelled by both expert and crowdsourced annotators who were specifically trained to identify features of moral outrage as defined by the social psychology and affective science literature (ref. 16 gives full details of the development and validation of DOC). We formed two groups of tweets to invite participants to participate in our DM study: moral outrage tweets and non-moral outrage tweets. We defined a moral outrage tweet as any tweet that DOC classified as having a 0.95 or higher probability of containing moral outrage expression and non-moral outrage tweets as those that DOC classified as having a 0.05 or lower probability of containing moral outrage expression. We collected high- and low-outrage tweets to obtain a wide range of outrage intensities to be used in the observer phase of the study (see below). This created a context that better matched how observers encounter emotional messages online (not all messages contain moral outrage expressions).

Next, we attempted to send DMs to users DOC identified as expressing moral outrage or non-moral outrage, only for users who had opted in to open direct messaging. By default, a Twitter user cannot receive a direct message from other users whom they do not follow. Thus, users who we were able to send DMs to were those who changed Twitter's default setting to allow open DMs. DMs were sent to users from our active research account: Yale Social Media Research Group (@yaleSMRG) which was openly described as an academic research account in the public profile. In the DM, users were asked whether they would be interested in participating in our research and were told that their responses would remain anonymous. Then, the DM asked users to rate how outraged and how happy they felt in the moment, specifically when they tweeted the message we identified, on a 1 (not at all) to 7 (very) Likert scale. We instructed them to think about how they were feeling specifically at the time of tweeting to ensure they were not reporting general feelings about the political topic. We displayed the tweet we wanted them to report on in the DM (see Supplementary Appendix A for the full DM message text). The DMs were sent 1–15 minutes after a user posted their tweet. The median time between the sending of a DM and a user response to the message was 106 minutes (study 1) and 25 minutes (study 2), showing the ability of our method to measure authors' emotions as close to the time of sending their message as possible.

Only direct messaging the users who had opted into open DMs helps to maintain user privacy but also comes with limits to user recruitment. Users with open DMs do not necessarily represent the average Twitter user. For instance, 11.73% of users in study 1 and 19.21% of study 2 users that we collected from the API had opted into open DMs. Of the users that we were ultimately able to DM, we observed a response rate of 6.41% in study 1 and 15.61% in study 2. We note that three authors were removed from the data in study 2, since they requested that we not use their responses in the study. Supplementary Section 1.5 gives analyses that rule out key concerns about selection bias and generalizability using our method.

In the observer phase, we recruited a separate set of politically partisan participants using the Prolific.ac recruitment platform. We aimed to have each tweet rated by a minimum of ten Democrats and ten Republicans. The rating consisted of making judgements about how outraged and how happy each tweet author was, using the same scale that the tweet authors used in the author phase. Toward this end, we recruited the following number of participants for the observer phase:

study 1, $N = 140$; study 2, $N = 189$; study 3, $N = 362$. We note that these sample sizes are slightly higher than what were preregistered because we oversampled to ensure each tweet had at least ten people given the constraints of our randomization scheme. In our randomization scheme, if a participant dropped from the study, it still led to a tweet getting counted as being seen once, which meant that we had to run more participants to account for participants dropping out. The following number of participants were removed in each study for failing to meet preregistered exclusion criteria (attention check about the topics of the tweets and not being Republican or Democrat): study 1, $N = 30$; study 2, $N = 24$; study 3, $N = 52$.

Each tweet was presented to participants in an anonymized form with the username and profile picture made blank. Each participant judged a stratified random selection of 30 tweets (15 were randomly drawn from those classified by DOC as outrage and 15 were randomly drawn from those classified by DOC as non-outrage).

To test whether observers were overperceiving the outrage/happiness in authors' tweets, we used a generalized linear model with judgements of outrage/happiness clustered by observer (every observer judged 15 authors/tweets) and author id and tweet id entered as random factors. Fixed effects estimated the difference in outrage values between authors and observers. For the main analyses, overperception of outrage and happiness were tested in two separate models but see Supplementary Section 1.2 for a model fit using both emotion judgements. We used the lmer function in the lme4 package in R v.3.4.3. All data and analysis scripts are available at https://osf.io/gtwsk/.

As a robustness check, we also computed the mean levels of the emotions perceived in each author tweet based on all the judgements of observers who viewed the tweet. For each tweet, we then tested the difference between the tweet author's self-report of outrage and happiness compared to the mean of all participants' judgements of how outraged/happy the tweet author was. To examine the difference, we conducted a Wilcoxon sign-ranked test for paired data as our data were non-normal.

To test the relationship between observers' overperception and their levels of daily political social media use, we asked participants to report whether they used social media daily to learn about politics and then to report the number of times they do so daily on a 0–100 sliding scale to capture even the most extreme social media users[48]. Some observers were not included in the analysis because they did not fully complete the social media use questionnaire which appeared at the end of the survey (study 1, $N = 23$; study 2, $N = 28$; study 3, $N = 59$). The mean level of overperception for each observer was determined as the mean discrepancy between the observers' judgements and the authors' self-reported outrage, for all tweets that an observer judged. Greater values indicated that an observer tended to perceive more outrage than reported by the authors of the tweets they viewed.

Preregistered experiment (study 4)

Studies 4 and 5 were approved by Yale University's IRB, approval no. 2000022385 and all participants consented to research by agreeing to an online informed consent form. We recruited 300 Democrats and 300 Republicans identified from Prolific.ac to participate in a study about 'making social judgements'. We report how we determined our sample size, all data exclusions, all manipulations and all measures in the study in our preregistration at https://osf.io/sxtah. After removing participants who were not politically partisan ($N = 27$) and then those who failed a comprehension check ($N = 52$), our final N was 523.

In the first part of the study, participants were randomly assigned to view one of two simulated Twitter newsfeeds. As described in the main text, the high-overperception newsfeed contained ten tweets for which the observers from studies 1 and 2 reported notably more outrage than the authors reported. The low-overperception newsfeed contained ten tweets for which the observers' judged outrage and authors' self-report were within 1 scale point on average. Crucially, these feeds were matched in the level of outrage self-reported by tweet authors (Fig. 4b). For both conditions, we chose tweets with lower outrage self-reported by authors to avoid ceiling effects of outrage judgements and to allow a wider range for bias to be measured. The mean of authors' self-reported outrage in the high-overperception condition ($M_{\text{republican}} = 1.20$; $M_{\text{democrat}} = 1.20$) and the low overperception ($M_{\text{republican}} = 1.30$; $M_{\text{democrat}} = 1.40$) was held at similar levels. However, the magnitude of overperception (mean observer outrage ratings − mean author outrage ratings) was much greater in the high-overperception condition ($M_{\text{republican}} = 4.25$; $M_{\text{democrat}} = 3.92$) compared to the low-overperception condition ($M_{\text{republican}} = 1.58$; $M_{\text{democrat}} = 1.36$). No social information (for example, a 'likes' count or a 'shares' count) was displayed on tweets.

In both conditions, participants only viewed political ingroup tweets based on their self-reported party identification (any participants who did not identify as Republican or Democrat ($N = 18$) were removed from the analysis). Participants were told that there would be a memory task at the end of the experiment to promote greater attention to tweet content in the newsfeeds (there was actually a memory task at the end that showed participants one tweet they saw and one tweet they did not see but the data were not analysed).

To measure how newsfeed condition affected judgements of collective outrage, we asked participants to judge the social network who composed the tweets in the newsfeed using the following question: 'Thinking about the social network on average, how morally outraged do you think members of this social network are?'. Participants responded using the same 1 (not at all outraged) to 7 (very outraged) Likert scale as authors and observers from studies 1–3. To test for differences in judgements of collective outrage between groups, we conducted an independent-samples t-test adjusting for unequal variances.

To test the predictions of the simple average versus weighted average models of social learning, a one-sample t-test was conducted to compare the collective outrage judgement mean against the mean of outrage perceived in each newsfeed message when viewed individually (a message's 'individually perceived outrage' value). To determine the individually perceived outrage values, each tweet was assigned a value, representing the mean outrage judged by all observers who viewed the tweet. For example, consider 'Tweet 1' which appeared in the high-overperception feed. Tweet 1 was viewed by ten observers in study 1 and so it would be assigned an outrage value calculated as the mean of all ten outrage judgements made by observers in study 1. Following this example, all the tweets in both newsfeed conditions had 1 individually perceived outrage value and the mean of those values represented the mean of outrage perceived in each newsfeed message when viewed individually. These means were then used as the comparison value in a one-sample t-test. For example, the mean collective outrage judgement in the high-overperception condition (across Democrats and Republicans) was 5.96. This value was compared against 4.09, which was the mean of all the individually perceived outrage values for messages appearing in the high-overperception newsfeed.

Preregistered experiment (study 5)

We recruited 600 Democrats and 600 Republicans to participate in a study about making social judgements. 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/mjftk. After removing participants who were not politically partisan ($N = 100$) and then those who failed a comprehension check ($N = 87$), our final $N = 1{,}013$.

Using the same manipulation as in study 4, participants were randomly assigned to either the high-overperception newsfeed or the low-overperception newsfeed (Fig. 4 and study 4). After viewing the newsfeed, participants were first asked to make judgements of outrage norms in a norm judgement task. In this task, we asked participants to view ten new tweets about the 2020 US election and Amy Coney Barrett confirmation hearings. For the ten new tweets, we manipulated

the extent to which they expressed moral outrage. There were five new high-outrage tweets that were from studies 1 and 2 (tweets that had not appeared in any newsfeeds in either condition), where high outrage was defined as an author self-reported outrage greater than a 4 on the 1 (not at all outraged) to 7 (very outraged) scale. We only used tweets that had low-overperception scores (no more than 1 scale point difference between author and observer ratings on the outrage scale). The other five new tweets were more neutral tweets about the same political topics (2020 US election and Amy Coney Barrett confirmation hearings).

Every participant viewed all ten tweets, viewed one tweet per trial and were asked the following question on each trial: 'How socially appropriate/inappropriate would it be for someone to post this tweet to the social media network that sent the messages you recently viewed?' Participants responded on a −3 (very socially inappropriate) to 3 (very socially appropriate) scale. Our dependent variable of interest was the difference score between the mean appropriateness ratings for the high-outrage tweets versus the neutral tweets, representing the appropriateness of outrage tweets relative to the appropriateness of non-outrage tweets.

After the norm judgement task, participants also make judgements about group affective polarization. To measure how condition affected perceptions of group affective polarization, we asked participants to judge how much the social network they viewed liked the political ingroup and outgroup using a 0–100 feeling thermometer scale, which is a standard measure of affective polarization[31]. Specifically, participants were asked, 'On average how do you think the people in this network feel about Republicans (Democrats)?'.

Finally, participants made judgements about ideological extremity. To measure how condition impacted perceptions of ideological extremity, we asked participants, 'What do you think the political ideology of the typical person in the network is?', on a −3 (extremely liberal) to 3 (extremely conservative) scale. Ideological extremity was defined as the absolute value of judgements such that greater values indicated greater judgements of ideological extremity for both parties.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

All de-identified data are available at https://osf.io/gtwsk/ and https://doi.org/10.17605/OSF.IO/GTWSK. Data may not be used for commercial purposes.

Code availability

All analysis scripts are available at https://osf.io/gtwsk/ and https://doi.org/10.17605/OSF.IO/GTWSK. Code may not be used for commercial purposes.

References

1. Rini, R. *Social Media Disinformation and the Security Threat to Democratic Legitimacy* (NATO Association of Canada, 2019); https://natoassociation.ca/wp-content/uploads/2019/10/NATO-publication-.pdf
2. Raz, J. *The Morality of Freedom* (Clarendon Press, 1986).
3. Tattersall, A. *Power in Coalition: Strategies for Strong Unions and Social Change* (Routledge, 2020).
4. Lazer, D. M. J. et al. The science of fake news. *Science* **359**, 1094–1096 (2018).
5. Starbird, K., Arif, A. & Wilson, T. Disinformation as collaborative work: surfacing the participatory nature of strategic information operations. *Proc. ACM Hum. Comput. Interact.* **3**, 1–127 (2019).
6. Bail, C. *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing* (Princeton Univ. Press, 2021).
7. Salerno, J. M. & Peter-Hagene, L. C. The interactive effect of anger and disgust on moral outrage and judgments. *Psychol. Sci.* **24**, 2069–2078 (2013).
8. Haidt, J. in *Handbook of Affective Sciences* (Davidson, R. J. et al.) 852–870 (Oxford Univ. Press, 2003).
9. Brady, W. J., Crockett, M. J. & Van Bavel, J. J. The MAD model of moral contagion: the role of motivation, attention, and design in the spread of moralized content online. *Perspect. Psychol. Sci.* **15**, 978–1010 (2020).
10. Spring, V. L., Cameron, C. D. & Cikara, M. The upside of outrage. *Trends Cogn. Sci.* **22**, 1067–1069 (2018).
11. Brady, W. J. & Crockett, M. J. How effective is online outrage? *Trends Cogn. Sci.* **23**, 79–80 (2019).
12. Lengel, R. H. & Daft, R. L. The selection of communication media as an executive skill. *Acad. Manag. Perspect.* **2**, 225–232 (1988).
13. Weisband, S. & Atwater, L. Evaluating self and others in electronic and face-to-face groups. *J. Appl. Psychol.* **84**, 632–639 (1999).
14. Byron, K. Carrying too heavy a load? The communication and miscommunication of emotion by email. *Acad. Manag. Rev.* **33**, 309–327 (2008).
15. Walther, J. B. & D'Addario, K. P. The impacts of emoticons on message interpretation in computer-mediated communication. *Soc. Sci. Comput. Rev.* **19**, 324–347 (2001).
16. Brady, W. J., McLoughlin, K., Doan, T. N. & Crockett, M. How social learning amplifies moral outrage expression in online social networks. *Sci. Adv.* **7**, eabe5641 (2021).
17. Brady, W.J. et al. Algorithm-mediated social learning in online social networks. Preprint at *OSF preprints*. https://doi.org/10.31219/osf.io/yw5ah (2023).
18. Brady, W. J. & Van Bavel, J. J. Social identity shapes antecedents and functional outcomes of moral emotion expression in online networks. Preprint at *OSF preprints* https://doi.org/10.31219/osf.io/dgt6u (2021).
19. Jordan, J. J. & Rand, D. G. Signaling when no one is watching: a reputation heuristics account of outrage and punishment in one-shot anonymous interactions. *J. Personal. Soc. Psychol.* **118**, 57 (2019).
20. Jordan, J. J., Hoffman, M., Bloom, P. & Rand, D. G. Third-party punishment as a costly signal of trustworthiness. *Nature* **530**, 473–476 (2016).
21. Crockett, M. J. Moral outrage in the digital age. *Nat. Hum. Behav.* **1**, 769–771 (2017).
22. Brady, W. J., Wills, J. A., Jost, J. T., Tucker, J. A. & Van Bavel, J. J. Emotion shapes the diffusion of moralized content in social networks. *Proc. Natl Acad. Sci. USA* **114**, 7313–7318 (2017).
23. Degroot, M. H. Reaching a consensus. *J. Am. Stat. Assoc.* **69**, 118–121 (1974).
24. Goldenberg, A., Weisz, E., Sweeny, T. D., Cikara, M. & Gross, J. J. The crowd-emotion-amplification effect. *Psychol. Sci.* **32**, 437–450 (2021).
25. McClain, C. *70% of U.S. Social Media Users Never or Rarely Post or Share about Political, Social Issues* (Pew Research Center, 2021); https://www.pewresearch.org/fact-tank/2021/05/04/70-of-u-s-social-media-users-never-or-rarely-post-or-share-about-political-social-issues/
26. Duggan, M. & Smith, A. *The Political Environment on Social Media* (Pew Research Center, 2016); https://www.pewresearch.org/internet/2016/10/25/the-political-environment-on-social-media/
27. Huszár, F. et al. Algorithmic amplification of politics on Twitter. *Proc. Natl Acad. Sci. USA* **119**, e2025334119 (2022).
28. Chakradhar, S. *More Internal Documents Show how Facebook's Algorithm Prioritized Anger and Posts that Triggered it* (Nieman Lab, 2021); https://www.niemanlab.org/2021/10/more-internal-documents-show-how-facebooks-algorithm-prioritized-anger-and-posts-that-triggered-it/

29. Chowdhury, R. Examining algorithmic amplification of political content on Twitter. *Twitter* https://blog.twitter.com/en_us/topics/company/2021/rml-politicalcontent (2021).
30. Shamir, J. & Shamir, M. Pluralistic ignorance across issues and over time: information cues and biases. *Public Opin. Q.* **61**, 227–260 (1997).
31. Prentice, D. A. & Miller, D. T. Pluralistic ignorance and alcohol use on campus: some consequences of misperceiving the social norm. *J. Pers. Soc. Psychol.* **64**, 243–256 (1993).
32. Lees, J. M. & Cikara, M. *Understanding and Combating Misperceived Polarization*. *Phil. Trans. R. Soc. B* **376**, 20200143 (2020).
33. Wilson, A. E., Parker, V. A. & Feinberg, M. Polarization in the contemporary political and media landscape. *Curr. Opin. Behav. Sci.* **34**, 223–228 (2020).
34. Levendusky, M. S. & Malhotra, N. (Mis)perceptions of partisan polarization in the American public. *Public Opin. Q.* **80**, 378–391 (2016).
35. Enders, A. M. & Armaly, M. T. The differential effects of actual and perceived polarization. *Polit. Behav.* **41**, 815–839 (2019).
36. Ronson, J. *So You've Been Publicly Shamed* (Penguin, 2015).
37. Peters, K. & Kashima, Y. A multimodal theory of affect diffusion. *Psychol. Bull.* **141**, 966–992 (2015).
38. Lerman, K., Yan, X. & Wu, X.-Z. The 'majority illusion' in social networks. *PLoS ONE* **11**, e0147617 (2016).
39. Alipourfard, N., Nettasinghe, B., Abeliuk, A., Krishnamurthy, V. & Lerman, K. Friendship paradox biases perceptions in directed networks. *Nat. Commun.* **11**, 707 (2020).
40. Schöne, J. P., Parkinson, B. & Goldenberg, A. Negativity spreads more than positivity on Twitter after both positive and negative political situations. *Affect. Sci.* **2**, 379–390 (2021).
41. Kuran, T. Preference falsification, policy continuity and collective conservatism. *Econ. J.* **97**, 642–665 (1987).
42. Noelle-Neumann, E. The spiral of silence a theory of public opinion. *J. Commun.* **24**, 43–51 (1974).
43. Hampton, K., Rainie, L., Dwyer, M., Shin, I. & Purcell, K. *Social Media and the 'Spiral of Silence'* (Pew Research Center, 2014); https://www.pewresearch.org/internet/2014/08/26/social-media-and-the-spiral-of-silence/
44. Lees, J. & Cikara, M. Inaccurate group meta-perceptions drive negative out-group attributions in competitive contexts. *Nat. Hum. Behav.* **4**, 279–286 (2020).
45. Kozyreva, A., Lorenz-Spreen, P., Hertwig, R., Lewandowsky, S. & Herzog, S. M. Public attitudes towards algorithmic personalization and use of personal data online: evidence from Germany, Great Britain, and the United States. *Humanit. Soc. Sci. Commun.* **8**, 117 (2021).
46. Rader, E. Examining user surprise as a symptom of algorithmic filtering. *Int. J. Hum. Comput. Stud.* **98**, 72–88 (2017).
47. Berlo, D. K. *The Process of Communication: An Introduction to Theory and Practice* (Holt, Rinehart & Winston, 1960).
48. Guess, A., Munger, K., Nagler, J. & Tucker, J. How accurate are survey responses on social media and politics? *Polit. Commun.* **36**, 241–258 (2019).

Acknowledgements

We thank members of the Crockett Lab for valuable feedback throughout the project. We also thank members of the Greene and Cushman Moral Psychology Research Lab, members of the Deghani Computational Social Science Laboratory and members of the Willer Polarization and Social Change Lab for feedback from a laboratory presentation of this work. We thank J. Lees for feedback on analyses. We thank A. Goolsbee who contributed to the construction of the observer-phase survey in studies 1–3. We thank A. Blevins for designing Figs. 1 and 4. This project was supported by the National Science Foundation, award no. 1808868 (awarded to W.J.B.) and the Democracy Fund, award no. R-201809-03031 (awarded to W.J.B. and M.J.C.). The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

W.J.B., K.L.M., M.G. and M.J.C. designed the research. W.J.B., K.L.M., M.T. and K.L. performed the research. W.J.B. analysed the data with input from M.J.C. W.J.B. wrote the paper with input from M.J.C. and all authors contributed to revisions.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41562-023-01582-0.

Correspondence and requests for materials should be addressed to William J. Brady or M. J. Crockett.

Peer review information *Nature Human Behaviour* thanks Robb Willer and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Peer reviewer reports are available.

nature portfolio

Corresponding author(s): William J. Brady, M.J. Crockett

Last updated by author(s): Feb 6, 2023

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	Data collection involved Python 3.8
Data analysis	Data analysis involved R 3.4.3, package lme4 v. 1.1-31, all data analysis scripts are available at https://osf.io/gtwsk/

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

All deidentified data are available at https://osf.io/gtwsk/

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☒ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	The studies are quantative including a field study to measure emotion responses and experiments that measure social judgments
Research sample	The sample for the field studies consists of U.S. Twitter users who discussed politics. Our experiments use partisan social media users recruited from Prolific.ac. We selected this sample because it is our population of interest (social media users) although it is not necessarily representative of all Americans or people in the world. We chose the Prolific platform because it tends to have better data quality (e.g., higher response rates and comprehension checks) than MTurk and enforces better pay standards. Study 4 age: M = 35.61, SD = 13.33; 55% Female; Study 5 age: M = 36.95, SD = 13.94, 52% Female.
Sampling strategy	We used convenience sampling based on identifying specific types of social media users. Sample sizes for the field studies was based on a feasibility, sample size for our experiments was based on a priori power analyses.
Data collection	Data was collected by following live Twitter conversations in the field studies, and by recruiting participants on Prolific.ac in the experiments. Experimenter was not blind to hypothesis, participants were blind to hypothesis. No one was present aside from the researcher and participants.
Timing	Part 1 of the first field study was conducted 7/28/20 - 8/3/20. Part 2 of the first field study was collected 8/12/20-8/18/20. The second field study was collected 10/22/20 - 10/30/20E. Study 3 was collected on 11/20/20. Study 4 was collected on 3/31/21 and Study 5 was collected 12/1/20.
Data exclusions	Participants were removed from each study based on pre-registered criteria (comprehension check, were politically partisan). The total participants removed across the 5 studies was N = 379
Non-participation	In Study 2, 3 participants requested for their responses not to be used in the study, so we removed those responses from the data.
Randomization	In the experiments participants were randomly assigned to high vs. low overperception groups. In the field studies, there were no experimental groups. Tweet authors were Twitter users selected through the sampling strategy described above. Observers were recruited via Prolific.ac via the strategy described above.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☐	☒ Human research participants
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging

Human research participants

Policy information about studies involving human research participants

Population characteristics	See above
Recruitment	In the field studies, participants were recruited based on their political behavior on social media. These users represent those who made be more politically active than the typical user, but in our case it was our target sample. We tested for several potential sources of select-selection bias including mean outrage levels, ideological extremity and overall twitter activity, but

	we did not find evidence of any systematic differences between users who responded to our DMs and those who did not, see SI Appendix Section 1.5. There are no obvious sources of self-selection bias in our recruitment of Study 4 and 5 participants other than Prolific participants tend to skew more liberal than the average US population. However, because we did not find any political partisan differences in our results this potential source of bias is not relevant.
Ethics oversight	Studies were approved by Yale University IRB, approval #2000022385 and #2000026899. In Studies 1-3, Observer phase participants consented to research by agreeing to an online informed consent form. Study 4 and 5 participants participants consented to research by agreeing to an online informed consent form

Note that full information on the approval of the study protocol must also be provided in the manuscript.