PNAS NEXUS

PNAS Nexus, 2025, 4, pgaf310
https://doi.org/10.1093/pnasnexus/pgaf310
Advance access publication 16 December 2025
Research Report

Americans overestimate how many social media users post harmful content

Angela Y. Lee [a,*,†], Eric Neumann [b,†], Jamil Zaki [b] and Jeffrey Hancock [a]

[a]Department of Communication, Stanford University, Palo Alto, CA 94305, USA
[b]Department of Psychology, Stanford University, Palo Alto, CA 94305, USA
*To whom correspondence should be addressed: Email: angela8@stanford.edu
†Co-first authorship.
Edited By Sacha Altay

Abstract

Americans can become more cynical about the state of society when they see harmful behavior online. Three studies of the American public ($n = 1{,}090$) revealed that they consistently and substantially overestimated how many social media users contribute to harmful behavior online. On average, they believed that 43% of all Reddit users have posted severely toxic comments and that 47% of all Facebook users have shared false news online. In reality, platform-level data shows that most of these forms of harmful content are produced by small but highly active groups of users (3–7%). This misperception was robust to different thresholds of harmful content classification. An experiment revealed that overestimating the proportion of social media users who post harmful content makes people feel more negative emotion, perceive the United States to be in greater moral decline, and cultivate distorted perceptions of what others want to see on social media. However, these effects can be mitigated through a targeted educational intervention that corrects this misperception. Together, our findings highlight a mechanism that helps explain how people's perceptions and interactions with social media may undermine social cohesion.

Keywords: toxicity, social media, hate, misperception, beliefs

Significance Statement

Americans can become excessively cynical about their fellow citizens when they mistakenly believe that many people post harmful content online, neglecting to realize that most such content is produced by a small but prolific group of social media users. This makes people feel negatively and believe the nation is in moral decline. Fortunately, this misperception can be corrected by teaching people that social media can overrepresent the views of vocal accounts that post disproportionately often.

Introduction

When US-Americans go on social media, how many of their fellow citizens do they expect to post harmful content? Researchers have investigated the actual prevalence of "hateful, aggressive, and disrespectful" content online. Less work has examined the perceived prevalence of such users. Such perceptions are crucial to behavior according to the research that does exist. When people perceive a community as engaging in harmful acts, they can be less likely to participate or come to act more harmful themselves (1, 2). In this paper, we contrast the perceived and actual prevalence of users who post harmful content.

In reality, very few online users post harmful content. While there is no doubt that harmful content exists online (3), most of it is produced by a small but highly active subset of users who post prolifically (4, 5). For instance, 1% of conflict-seeking Reddit communities produced 74% of all conflict across the platform (6). And 60% of hateful speech on Twitter came from a small right-wing community (7). These are part of a broader pattern of online content production, which follows a power law: the majority of content is produced by a small, vocal minority at the heavy tail of the distribution (8, 9).

But people may not realize that most harmful online content comes from such a small minority of users. Past research provides multiple potential explanations for why people might instead overestimate how many social media users produce harmful content. People generally overestimate the prevalence of small demographic groups (10) and thus could overestimate the prevalence of the small group of harmful users as well. People might also pay attention to the amount of harmful content they see, without tracing who generates it, and thus mistake a lot of harmful content for a lot of harmful users (2). Next, negative content tends to get amplified by social media algorithms (11), which could make its producers look more

Competing Interest: The authors declare no competing interests.
Received: February 14, 2025. **Accepted:** July 17, 2025

prevalent than they are. Finally, people generally pay more attention to negative than positive content and may remember it more (12).

If people do overestimate the prevalence of harmful users, this misperception could foster excessive cynicism about their fellow citizens. When people believe that many of their fellow Americans are posting harmful content, they may develop more negative views of society and perceive greater moral decline than actually exists (13). Understanding these potential dynamics is crucial for addressing how social media interactions may shape people's fundamental beliefs about human nature and the character of their communities—online and offline.

In this paper, we investigate Americans' beliefs about how many social media users contribute to harmful content in three national studies and examine the consequences of such beliefs ($n = 1{,}090$). By comparing public perceptions against platform-level data, we demonstrate that people substantially overestimate how many users post severely toxic content and share false news (studies 1 and 2). This misperception affects how participants feel, how they think about the moral state of the nation, and what online content they think their fellow citizens value (study 3).

Results

Study 1 ($n = 295$, 48.2% White, 51% male, $M_{age} = 43$) compared participants' perceived prevalence of Reddit users posting harmful content to the actual prevalence of such users, as documented in one recent paper (14). This paper looked at one specific form of harmful content (i.e. content involving "toxic language") as measured by Google's Perspective API. Content was considered toxic when the API thought 90% of humans would see it as "hateful, aggressive, and disrespectful." After learning about this methodology, participants estimated how many Reddit users were found to have posted toxic content on Reddit (see Methods for details and Supplementary Material for the exact wording used). They substantially overestimated the actual finding. They believed that over a third of all active accounts had posted such content at least once over the study's 18-month period (mean = 38.1%, median = 30%). In fact, only 3% of active Reddit accounts had done so (14), $V = 43{,}382$, $P < 0.001$. These analyses replicate in studies 2 and 3, and meta-analytic results are displayed in Fig. 1a. Even if the paper had used a much more lenient definition of toxicity, participants still would have overestimated the findings (Table 1). There were no significant associations between the time people spent on Reddit ($\beta = 0.20$, $P = 0.38$) or social media in general ($\beta = 0.10$, $P = 0.38$) and their estimation accuracy.

This misperception was not limited to Reddit or toxic language. We next explored if participants would also overestimate how many social media users share false news, another type of content which can spread harm (16). It is important to note that hate speech and false news (i.e. misinformation or disinformation) constitute two distinct forms of social media content, with research indicating that people hold different motivations for posting each, to different effects (17, 18). However, we chose to examine these phenomena together because both (i) are posted by small groups of prolific accounts (4), (ii) are often perceived as degrading the quality of online discourse by both the public (19) and experts (20), and (iii) allow us to test whether misperceptions of harmful content prevalence represent a general phenomenon that extends across different types of problematic online behavior. After reading about another study ((15); see supplements for the exact wording we used), people overestimated that almost half of all Facebook users had shared false news online (46.8%, median = 50%), compared with the actual finding of 8.5%, $V = 43{,}130$, $P < 0.001$ (15). They also overestimated that a third of users were *super-sharers* who had shared 10 or more false news articles (36.5%, median = 32%), compared with the actual finding of <0.5%, a roughly 100-fold overestimation, $V = 43{,}660$, $P < 0.001$ (Fig. 1b).

People also failed to account for the outsized digital footprint of the few prolific accounts. Kumar et al. (14) found 3.1% of Reddit accounts posted toxic content, yet generated 33.3% of all content on the platform, with over 559 million comments. This equates to a 1:11 ratio. In contrast, participants believed that 38.3% of all Reddit accounts had posted toxic content and that this group was responsible for 38.1% of all content, toxic or nontoxic (Fig. 1c). Even when using within-person analyses, this results in a ratio of 1:2.6, a significant misestimation ($V = 1{,}838$, $P < 0.001$). This indicates that participants consistently underestimated the disproportionate contribution of toxic accounts to Reddit's content—underscoring the vast gap between reality and perception.

These findings reveal a striking pattern in how people misunderstand online toxicity. Participants drastically overestimated *how many users* posted toxic comments on social media—believing it was 38% when it is actually 3% (a 13-fold overestimation). However, they were nearly accurate about *how much total content* this set of users produces—estimating 38% when it is actually 33% (a 1.15-fold overestimation). This suggests people may encounter toxic content at roughly the expected volume but attribute it to far more widespread participation than actually occurs. Rather than recognizing a small group of highly active accounts, people appear to imagine toxic behavior as broadly distributed across the user base.

These misperceptions did not simply reflect a misunderstanding of how harmful behavior was defined in the research. In study 2 ($n = 185$, 71.3% White, 49.8% male, $M_{age} = 46$, preregistered), participants completed a signal detection task where they saw 20 sample comments from the above Reddit toxic language study (14). Then they indicated which comments they believed were classified as toxic in the paper. Curiously, participants were highly accurate ($M_{d'} = 2.45$, $SD_{d'} = 1.57$) and were not biased towards under- or overestimating what content the algorithm would consider toxic ($M_{criterion} = 0.02$, $SD_{criterion} = 1.33$; Fig. 1d). Yet, they once more overestimated how many accounts posted toxic content, believing that 45% had done so (median = 45%), $V = 17{,}157$, $P < 0.001$. In fact, out of the 185 participants, merely eight *did not overestimate* how many Reddit users had posted toxic content. People understood how toxic content was defined, and nonetheless believed it was coming from many more users than it truly is.

This misperception had consequences. Study 3 ($n = 611$, 65.9% White, 49.5% male, $M_{age} = 45$, preregistered) was an experiment that randomly assigned participants to one of two conditions. In the misperception correction condition, participants once more estimated the Reddit toxicity findings and then learned about the actual findings as well as findings from two related studies (6, 15). In the control condition, participants simply read about the history of Reddit. Participants in the misperception correction condition felt more positive ($\beta = 0.59$, SE = 0.07, $P < 0.001$; for a more detailed description of all outcome variables in this study, see Supplementary Material). They also came to believe the character of their fellow US citizens was in less moral decline (13) than those in the control condition, $\beta = -0.23$, SE = 0.10, $P = 0.02$. This is noteworthy given how short and context-specific our correction condition is and how persistent the belief in moral decline is, which replicates across at least 60 countries. Next, we found a novel form of "pluralistic ignorance," where people can underestimate how much their in-group members

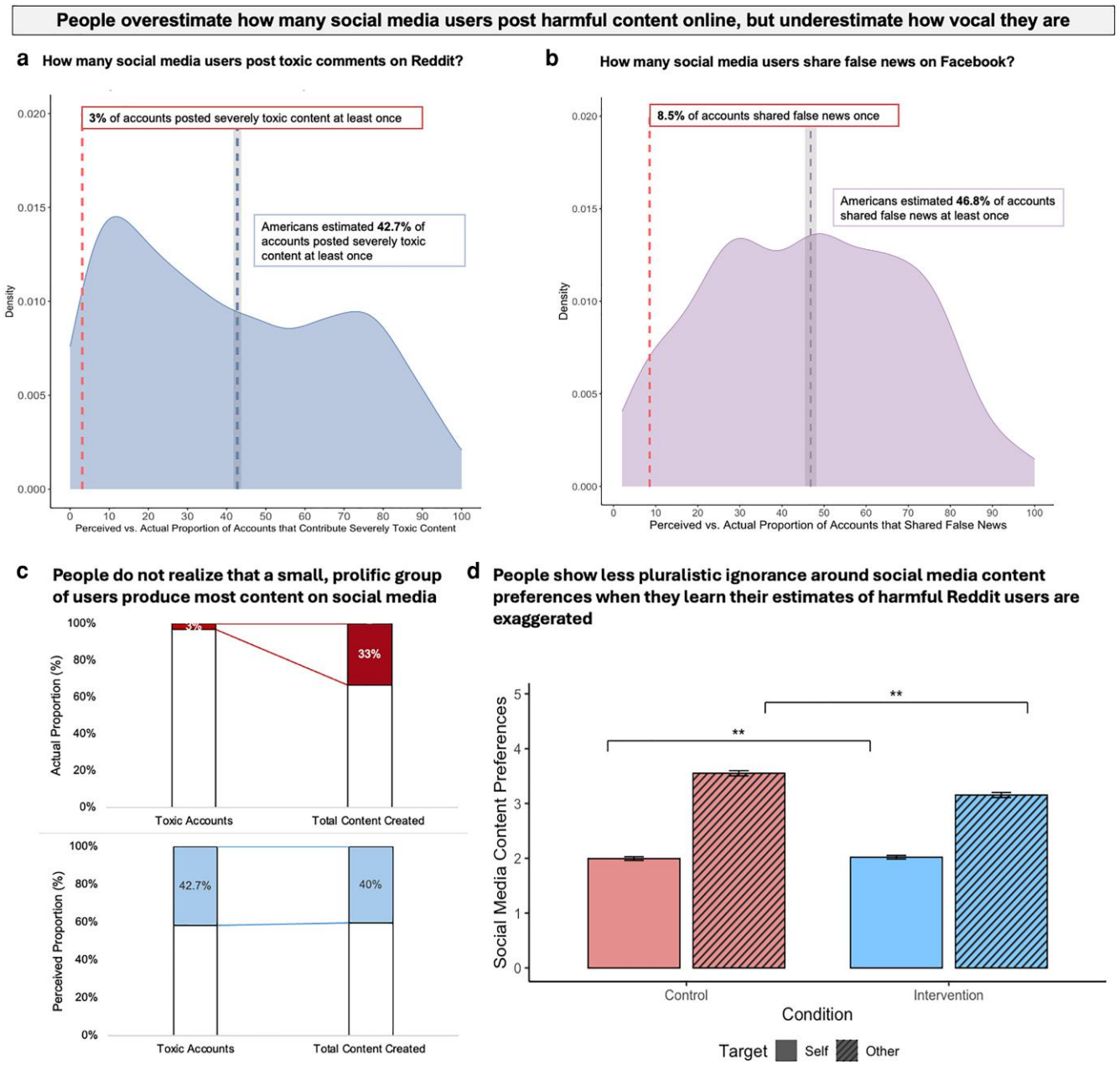

Fig. 1. People overestimate how many social media users post harmful content online, but underestimate how vocal they are. a) Density plot indicating that people overestimated how many Reddit accounts posted severely toxic content (data from (6)). Dashed red lines indicate the percentage of accounts that posted severely toxic content at least once. Blue line indicates Americans' estimate of accounts that posted severely toxic content, meta-analyzed across studies 1–3, shading indicates SE. b) Density plot indicating that people overestimated how many Facebook accounts shared false news (data from (15)). Dashed red line indicates the percentage of accounts that shared false news at least once. Blue line indicates Americans' estimates (study 1), shading indicates SE. c) Visualization of actual and perceived contribution ratios of toxic accounts on Reddit. Red plot indicates that 3% of all active accounts on Reddit produced toxic content and contributed 33% of all content on Reddit (data from (6)). Blue plot indicates that meta-analyzed across our three studies, participants believed 42.7% of all active accounts on Reddit produced toxic content and contributed 40% of all content on Reddit. d) Graph showing differences in pluralistic ignorance around social media content perceptions between the correction condition (blue) and the control condition (red). Participants that learned their estimates of Reddit users that posted toxic comments were exaggerated demonstrated reduced self-other differences in beliefs about the kinds of content they believed should go viral online (nonshaded bars), and the kinds of content other Americans believed should go viral online (shaded bars). *Indicates P-value < 0.05, **indicates P-value < 0.01, and ***indicates P-value < 0.001.

share their beliefs (21). Specifically, we found that our national sample of US participants underestimated how much their average fellow citizen shared their desire for less harmful online content, $V = 40{,}738$, $P = 0.003$ (inspired by (19)). Critically, participants in the correction condition were less likely to succumb to this pluralistic ignorance relative to the control, i.e. were more likely to understand that others often do not desire harmful online content, $\beta = -0.95$, SE = 0.34, $P = 0.005$. Finally, we tested whether participants might generalize even more from their corrected misperception and change their global beliefs about other people. Finally, and contrary to our hypotheses, there were no significant differences in participant's cynicism ($\beta = -0.03$, SE = 0.09, $P = 0.76$) and generalized trust in human nature ($\beta = -0.02$, SE = 0.12, $P = 0.88$) between the two conditions.

Table 1. Magnitude of misperceptions between perceived harmful social media behaviors and actual harmful social media behaviors.

Proportion of Reddit accounts that posted at least one severely toxic comment				
Perceived prevalence (meta-analyzed from studies 1 to 3)	**Misperception (magnitude of difference)**	**Actual prevalence (platform-level data from Reddit)**		
		Toxicity threshold	**Number of accounts classified as toxic**	**Percentage of accounts classified as toxic**
41.46%	+38.4% V = 312,839***	0.90	959,007	3.1
	+27.6% V = 287,523***	0.80	4,758,998	13.9
	+20.9% V = 259,220***	0.70	7,044,549	20.6
	+12.1% V = 222,633***	0.60	9,709,756	29.4
	+8.4% V = 200,764***	0.50	11,306,010	33.1

Comparisons of lay beliefs about the proportion of Reddit accounts that had posted at least one severely toxic comment, relative to platform-level data assessing the number of accounts that posted severely toxic comments with five different thresholds (6). We re-analyzed the original dataset from Kumar et al. (14). To identify the number of active Reddit accounts that posted at least one severely toxic content when different thresholds for severely toxic content were used. The perspective API produces a score from 0 to 1.00, indicating the percentage of people, out of ten, that would view a given comment as severely toxic. While the original paper used the recommended threshold where nine out of 10 people (0.90) would view a given comment as severely toxic, we calculated the number of accounts that would be classified as contributing to severely toxic content for the thresholds of 0.5, 0.6, 0.7, and 0.8. The significance of the mean difference between lay beliefs and platform-level data was assessed with Wilcoxon signed-rank tests.

Discussion

Our results reveal people do not realize that most harmful content on social media is produced by a small, prolific group of users. Instead, they believe that the amount of *harmful content* on social media is the result of many *users* participating in harmful behaviors. Across three studies, participants vastly overestimated how many Reddit users posted toxic content, 13-fold (meta-analyzed $M = 41.46\%$ vs. 3.1% in reality). Overestimates were not confined to this domain of harmful online content (study 1) and were not due to misunderstood definitions of toxicity (study 2). This indicates many people mistake an extremely vocal minority for a somewhat vocal majority, failing to realize that most social media users never post harmful content online. This may be part of a more general cognitive phenomenon—individuals' tendency to hedge estimates of proportions towards more central prior beliefs (21), which causes them to generally overestimate minority groups—that occurs in social media as well as in other contexts. More specifically, this misperception may be driven by negative media coverage of social media that makes the harmful actions of small online minorities to be particularly salient (22). Finally, social media may be a particularly poor representation of the prevalence of toxic sentiment in society because it is easier for people to express socially undesirable views while anonymous (2, 23).

This particular, widely held misperception about social media toxicity can have serious consequences, leaving participants feeling negatively, with higher pessimism about the moral state of their fellow citizens, and with more pluralistic ignorance about how much others desire harmful online content (study 3). Witnessing outpourings of digital conflict can make people feel that society itself is becoming more toxic than ever, *when they mistakenly believe that this toxicity reflects the public* (2). Our work provides empirical support for the theory that social media harms social cohesion, albeit through a novel perceptual mechanism. Our results demonstrate that people overestimate the prevalence of harmful people on social media—which in turn may make them feel more negatively about their fellow citizens. Fortunately, this misperception can be corrected through targeted correction. If social media platforms are to remain a part of modern society, people should recognize that the opinions they see are not representative of public opinion.

Future research could examine specific *behaviors* that result from these misperceptions and the particular features (i.e. anonymous posting, pseudonymous posting) on the effects of misperceptions on individual feelings. Finally, while we aimed to show that misperceptions exist for online harmful behavior, more research is needed to understand if misperceptions are more, less, or equally pronounced for, e.g. offline harmful behavior or unharmful online behavior.

Materials and methods

From June 2023 to April 2024, we conducted three surveys of US-American adults via CloudResearch Connect, matched to national quotas of age, gender, race, and ethnicity from the 2020 United States Census. Procedures were approved by the Stanford University Institutional Review Board and all participants provided informed consent and were compensated financially for their time.

Study 1 asked participants to read about two research studies that identified how many Reddit accounts had posted toxic content (14) and how many Facebook users had posted false news on the platform (15). Participants read detailed descriptions of the definitions and methodology precisely as used in the research. Participants then provided their estimates regarding how many social media users produced such content (see Supplementary Material for full stimuli). Comparisons of lay beliefs against study findings were conducted with Wilcoxon signed-rank tests, given the non-normal distributions of participant estimates. To test whether people were more accurate depending on their Reddit or social media usage, we fit a series of linear models that regressed each of those usage types onto an accuracy variable, controlling for age and gender. The accuracy variable was the difference between participants' estimates and study findings.

Study 2 participants completed a signal detection task focusing on their comprehension of the Google Perspective API classifier used to detect toxic language in the research (14). They learned that the "severe toxicity" classifier was trained on trained coders' manual annotations to determine if comments were "hateful, aggressive, disrespectful, or otherwise likely to make a user leave a discussion or give up on sharing their perspective." They viewed 20 comments from the original Reddit dataset (14), half of which were classified as severely toxic and half which were not. Then, they were asked to identify the comments the classifier would

code as severely toxic. Sensitivity (d') and criterion were calculated for each individual. A positive criterion indicated that participants had a stricter understanding of toxicity than the classifier, and a negative criterion indicated a looser understanding.

Study 3 was an experiment where participants in the misperception correction condition read a short two-paragraph text about the abovementioned findings. The text claimed that these findings reflect a "general trend scientists have discovered recently: most people never share toxic online content." Participants in the control condition entered a time-matched control condition, where they learned about how Reddit was founded. This text did not contain any information about toxicity. Participants in both conditions completed measures of social media use, cynicism, generalized trust, perceptions of moral decline, and beliefs about the kinds of content that should go viral on social media (see Supplementary Material for all measures). For each dependent variable, linear regression analyses were conducted with condition as the independent variable and participants' age, gender, and social media use as covariates.

Acknowledgments

The authors thank their reviewers for improving this manuscript. In addition, they also thank the Stanford Social Media Lab, including Sunny X. Liu, Ryan C. Moore, Ross Dahlke, Fangjing Tu, Harry Yaojan Yan, Will Schulz, and Ronald Robertson for providing feedback on this manuscript. A.Y.L. is supported by the Mark & Mary Stevens Stanford Interdisciplinary Graduate Research Fellowship and the Stanford Social Impact Labs.

Supplementary Material

Supplementary material is available at *PNAS Nexus* online.

Funding

No funding was received for this research.

Author Contributions

Angela Y. Lee (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing—original draft, Writing—review & editing), Eric Neumann (Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Visualization, Writing—original draft, Writing—review & editing), Jamil Zaki (Methodology, Resources, Writing—review & editing), and Jeffrey Hancock (Methodology, Resources, Writing—review & editing).

Preprints

This manuscript was posted as a preprint: https://osf.io/preprints/psyarxiv/f8y62_v1.

Data Availability

Anonymized data and code scripts are available on OSF (https://osf.io/g7u6k/?view_only=5704547f08cf44329c5c3fd3d1722746).

References

1 Matias JN. 2019. Preventing harassment and increasing group participation through social norms in 2,190 online science discussions. *Proc Natl Acad Sci*. 116(20):9785–9789.
2 Robertson CE, Del Rosario KS, Van Bavel JJ. 2024. Inside the funhouse mirror factory: how social media distorts perceptions of norms. *Curr Opin Psychol*. 60:101918.
3 Avalle M, *et al.* 2024. Persistent interaction patterns across social media platforms and over time. *Nature*. 628(8008):582–589.
4 Budak C, Nyhan B, Rothschild DM, Thorson E, Watts DJ. 2024. Misunderstanding the harms of online misinformation. *Nature*. 630(8015):45–53.
5 Mamakos M, Finkel EJ. 2023. The social media discourse of engaged partisans is toxic even when politics are irrelevant. *PNAS Nexus*. 2(10):pgad325.
6 Kumar S, Hamilton WL, Leskovec J, Jurafsky D. Community interaction and conflict on the web. In *Proceedings of the 2018 World Wide Web Conference* (WWW '18). International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, CHE, 2018. p. 933–943. https://doi.org/10.1145/3178876.3186141.
7 Evkoski B, Pelicon A, Mozetič I, Ljubešić N, Kralj Novak P. 2022. Retweet communities reveal the main sources of hate speech. *PLoS One*. 17(3):e0265602.
8 Mitzenmacher M. 2004. A brief history of generative models for power law and lognormal distributions. *Internet Math*. 1(2):226–251.
9 Barabasi AL. 2005. The origin of bursts and heavy tails in human dynamics. *Nature*. 435(7039):207–211.
10 Landy D, Guay B, Marghetis T. 2018. Bias and ignorance in demographic perception. *Psychon Bull Rev*. 25:1606–1618.
11 Brady WJ, Jackson JC, Lindström B, Crockett MJ. 2023. Algorithm-mediated social learning in online social networks. *Trends Cogn Sci*. 27(10):947–960.
12 Rozin P, Royzman EB. 2001. Negativity bias, negativity dominance, and contagion. *Pers Soc Psychol Rev*. 5(4):296–320.
13 Mastroianni AM, Gilbert DT. 2023. The illusion of moral decline. *Nature*. 618(7966):782–789.
14 Kumar D, Hancock J, Thomas K, Durumeric Z. Understanding the behaviors of toxic accounts on Reddit. In *Proceedings of the ACM Web Conference 2023 (WWW '23)*. Association for Computing Machinery, New York, NY, USA, 2023. p. 2797–2807.
15 Guess A, Nagler J, Tucker J. 2019. Less than you think: prevalence and predictors of fake news dissemination on Facebook. *Sci Adv*. 5(1):eaau4586.
16 Grinberg N, Joseph K, Friedland L, Swire-Thompson B, Lazer D. 2019. Fake news on Twitter during the 2016 US presidential election. *Science*. 363(6425):374–378.
17 Mao Y, Xu T, Kim KJ. 2023. Motivations for proactive and reactive trolling on social media: developing and validating a four-factor model. *Soc Media Soc*. 9(4):20563051231203682.
18 Melchior C, Oliveira M. 2024. A systematic literature review of the motivations to share fake news on social media platforms and how to fight them. *New Media Soc*. 26(2):1127–1150.
19 Rathje S, Robertson C, Brady WJ, Van Bavel JJ. 2024. People think that social media platforms do (but should not) amplify divisive content. *Perspect Psychol Sci*. 19(5):781–795.
20 Stanford Youth Safety and Digital Well-Being Report. 2025. *Center for Digital Health*. https://cdh.stanford.edu/sites/g/files/sbiybj29486/files/media/file/youth_safety_and_digital_wellbeing_report_2025.pdf
21 Guay B, Marghetis T, Wong C, Landy D. 2025. Quirks of cognition explain why we dramatically overestimate the size of minority groups. *Proc Natl Acad Sci*. 122(14):e2413064122.

22 Lyons BA, Merola V, Reifler J. How bad is the fake news problem—Stephan Lewandowsky. *The Psychology of Fake News*. London: Routledge, 2020. p. 11–26.

23 Kim JW, Guess A, Nyhan B, Reifler J. 2021. The distorting prism of social media: how self-selection and exposure to incivility fuel online comment toxicity. *J Commun*. 71(6):922–946.