Commentary

Future Challenges for Online, Crowdsourced Content Moderation: Evidence from Twitter's Community Notes

Valerie Wirtschafter and Sharanya Majumder

1 Introduction

In recent years, the content moderation policies of social media platforms have garnered significant attention and scrutiny (Gallo and Cho 2021; Kozyreva et al. 2023). Faced with the challenge of identifying and removing harmful content and curbing the spread of misinformation and disinformation, companies have invested heavily in a range of approaches (Klonick 2017). These approaches often leverage a mixture of automated detection tools and paid content moderation teams. The use of volunteers—culled from a platform's community of users—is also becoming an increasingly prominent approach, particularly as social media companies seek to cut costs, increase revenue, and deflect responsibility for potentially controversial moderation decisions (Allen et al. 2021). In some instances, these community-based approaches have proven to be fairly cost-effective and comprehensive. For example, it is estimated that on average volunteer moderation teams across Reddit do over 450 hours of work a day, amounting to around $3.4 million in free labor a year (Stokel-Walker 2022). However, recent research has also found that crowdsourced fact-checking may have limited efficacy, particularly among individuals with low political knowledge (Godel et al. 2021).

Following leadership changes in late 2022, Twitter has leaned into a "wisdom of the crowd" approach as the future of content moderation across the platform. Operationalized through a program called "Community Notes," the platform's new owner Elon Musk has highlighted the feature, which relies on the contributions of users to provide context to tweets, as a "gamechanger for improving accuracy" and has touted its potential "powerful impact on falsehoods" as a means for establishing a baseline truth.[1] Prior to this, Twitter relied primarily on a centralized approach that made use of both automated systems and human reviewers (Delcker 2022). The platform designed the crowdsourced solution as a complement to these traditional practices. After piloting the program for over a year, Twitter expanded the visibility of Community Notes contributions to everyone in the United States on October 6, 2022, and internationally on December 10, 2022.

Despite its rapid rollout across the platform, the efficacy of this crowdsourced content moderation approach at scale remains an open question. Although declining ad revenue

[1] Elon Musk (@elonmusk). Twitter post: *Community Notes is a gamechanger for improving accuracy on Twitter!*. November 10, 2022. URL: https://twitter.com/elonmusk/status/1590773586616545280; Elon Musk (@elonmusk). Twitter post: *When @CommunityNotes rolls out worldwide, it will have a powerful impact on falsehoods*. November 10, 2022. URL: https://twitter.com/elonmusk/status/1591098584128434176; Elon Musk (@elonmusk). Twitter post: *[In response to a tweet asking how to find a baseline truth:] Community Notes aka @birdwatch*. November 11, 2022. URL: https://twitter.com/elonmusk/status/1589454170909597696.

doi:10.54501/jots.v2i1.139

and usage across Twitter have dented the platform's appeal, governments, the private sector, media outlets, and other influential figures continue to rely on it regularly to reach audiences (Harwell 2022; Dang 2022; Peoples, Izaguirre, and Licon 2023). Furthermore, due to the open-sourced and transparent nature of the Community Notes process, existing and future social networking platforms aiming to bolster their content moderation practices with "crowdsourced" solutions might seek to emulate the program in some capacity as well. As a result, understanding the evolving nature of this program and its ensuing impacts is important for broader content moderation conversations.

To provide that understanding, we recently carried out a comprehensive analysis of Community Notes. Drawing from 52,000 Community Notes and 1.3 million ratings from 27,000 contributors, our analysis paints a mixed picture of the program's efficacy and points to a number of clear challenges with crowdsourced content moderation solutions more broadly. Encouragingly, we found that as the Community Notes program has expanded, the overall balance and quantity of notes has improved. Although just 20% of Community Notes contributors have written "helpful" notes, the number of "helpful" notes attached to tweets as additional context has nearly tripled since the program's expansion, from 4% to approximately 12% of all written notes.[2] Prior to the expansion of the Community Notes program, users estimated to be more liberal received more notes attached to their content than users estimated to be more conservative, but with the expansion of the program, increased visibility of Community Notes, and addition of new contributors, that partisan gap has also shrunk rapidly.

Perhaps more concerning given recent advances in generative artificial intelligence, we found that among the tweets with additional context attached to them that were still online as of mid-March 2023, almost one out of every five contained doctored or misleading images, videos, and quotes. Additionally, at least at this time, Community Notes do not seem to impact the subsequent behavior of Twitter users. When compared to prior engagement levels, note receivers do not experience an overall decline in engagement on tweets shared after their post received additional context from Community Notes contributors. Twitter users are also no less likely to delete tweets tagged with a "helpful" Community Note than those who do not receive a "helpful" note attached to their tweet, but have one in adjudication.[3]

Given the polarized nature of content moderation practices and their associated costs, social media sites seem poised to lean on their user bases to assess contentious content moving forward. Although crowdsourced content moderation practices like Community Notes show promise in small samples, the success of these "wisdom of the crowd" approaches moving forward will depend on their ability to scale rapidly and effectively to address a range of ongoing and emerging challenges. As crowdsourced content moderation solutions continue to evolve, their reach will need to increase rapidly to match the speed of content generation—a challenge given current features of the program, including the ability for users to write notes about content that is nearly a decade old. More broadly, crowdsourced solutions will have to contend with two countervailing trends: (1) an increasingly polarized information ecosystem, where crowdsourced consensus will be difficult to come by for critically important but politically charged issues, such as elections; and (2) the growing difficulty of identifying content produced by generative AI tools, such as Dall-E 2, ChatGPT, and Midjourney, which are now capable of generating convincing disinformation devoid of many prior markers of fabricated content, including linguistic errors. These challenges, which directly impact the efficacy of crowdsourced

[2] A "helpful" note is a note written by a Community Notes contributor that is rated "helpful" by enough contributors of "different perspectives" that it is ultimately attached as additional context to the tweet.

[3] A "note in adjudication" describes a note written by a Community Notes contributor that has been drafted but not rated "helpful" by enough contributors of "different perspectives." These notes are *not* attached publicly as additional context to the tweet.

content moderation, are likely to become even more acute in the coming years.

2 What is Crowdsourced Content Moderation?

First coined in 2006, the term "crowdsourcing" describes a process by which the collective knowledge of the masses is harnessed to solve concrete tasks, making their aggregate output more reliable and trustworthy (Howe 2006). Today, the process of tapping into "the wisdom of crowds" through crowdsourcing is common, with applications ranging from community monitoring to fundraising (Surowiecki 2004). It is unsurprising, then, that these efforts have also extended to moderating content online.

Over the past two decade, many prominent companies have implemented crowdsourced content moderation solutions. YouTube, for example, has dedicated teams to take down content that violates Community Guidelines and a "priority flagger" program, where individual users with a high accuracy rate for flagging content in violation of these guidelines have their flags prioritized for review (Google 2023). Reddit enlists the help of volunteer moderators to enforce overarching content policies and create additional ones for the active subreddits they manage (Renfro 2022). Community upvoting and downvoting by all users is designed to boost or decrease the visibility of these contributions (Singh 2019). Wikipedia's self-governing body of 43 million registered users, in combination with machine learning tools, helps protect the internet's arguably "most reliable source" from bad faith actors and coordinated disinformation campaigns (Stuart 2021; Borak 2022). On Wikipedia, anyone can contribute, though some "administrators" who have years of high-quality editing history can adjudicate disputes between users with less "status" and sanction users who violate Wikipedia's guidelines (Goldberg, Lo, and Novotny 2020; Wikipedia, n.d.).

Twitter's Community Notes represents the latest prominent crowdsourced solution rolled out across a major social media platform (Twitter 2023a, 2022). Like Wikipedia and Reddit's approaches to content moderation, Community Notes relies on the contributions and consensus of its community to assess content posted on the platform. These contributors provide input to Twitter by both writing and rating notes. Initially, contributors are only able to rate notes but not write them. As users rate the contributions of others, they boost their Rating Impact score, which tracks their role in helping a note receive a "helpful" or "unhelpful" status. Once a contributor unlocks the ability to write notes, other users can then rate their contributions on any tweet posted to Twitter.

Unlike many other crowdsourced solutions, the locked status of "helpful" or "unhelpful" is not achieved by majority rule. Instead, it is based on whether or not there is agreement on the note's helpfulness by contributors of "different perspectives." Contributors are thought of as having "different perspectives" based on how they have rated notes in the past and, specifically, whether or not they have disagreed on prior note ratings (Twitter 2023b). After two weeks, a note can no longer receive reviews from contributors, and its status is locked by Twitter as either "helpful," "unhelpful," or "needs more reviews" (Twitter 2023d).[4]

[4] Notes locked with the status "needs more reviews" never reached a definitive status of "helpful" or "unhelpful" due to a lack of consensus across "different perspectives."

3 How Effective is Community Notes in Moderating Content across Twitter?

Past research focused on Community Notes has for the most part assessed the pilot phase of the program, before it was available to all US-based users. This research has primarily explored the topics covered by contributors in their notes (Faife 2022) and identified political partisanship as a major driver of note writing and reviewing (Allen, Martel, and Rand 2022). Further research during the pilot phase revealed that only 5% of all notes written ended up being rated "helpful" and that more than half of the notes failed to include a single external source (Mahadevan 2021). Past research also suggests that Community Notes can improve the quality of Twitter content by reducing the prevalence of extreme sentiments (Borwankar, Zheng, and Kannan 2022).

To assess the implementation and impact of Community Notes both prior to and after the program's expansion, we draw on Community Notes data from the start of the program through February 24, 2023.[5] We also collected the tweets referenced in the Community Notes dataset. For tweets we were unable to collect using the unique tweet ID, we assume that the tweet was either deleted by the author or removed by Twitter.[6] For users with a note locked as "helpful" on one of their tweets, we also pulled their Twitter posts 10 days prior to receiving the note, the day they received the note, and nine days after.[7] To assess the partisan leanings of the 733 unique Twitter users who still had a tweet posted online with a Community Note attached to it on February 24, 2023, we pulled their following networks and estimated their partisan ideologies using a method common to computational social science research (Barberá 2015).[8] In addition, we collected follower information for the users who had a tweet online with a Community Note from Social Blade, an online analytics tool that tracks a range of data points from major social media platforms. This data includes the number of total followers the accounts had in March 2023, and the change in followers the week before, the week of, and the week after receiving a Community Note.[9]

Based on this data, we find that most Community Notes contributors have never written a "helpful" note. Out of the nearly 7,000 unique contributors who have written notes, more than 5,500 (≈80%) have never written a note locked as "helpful." Of those who have written a note locked as "helpful," 92% have done so four or fewer times. Only five users have written 50 or more "helpful" notes.

Furthermore, of the 52,000 notes in the dataset, only around 3,400 (or 7%) of all written notes have reached a locked status of "helpful." But the percentage that have done so since the program's expansion has improved over time. Prior to the program's expansion, just 4% of all notes reached "helpful" status. After the expansion, the percentage of

[5] Twitter makes all data regarding Community Notes contributions publicly available, including all notes that have been written about tweets, the ratings of these notes, and the status of these notes. The data can be downloaded here: https://github.com/twitter/communitynotes.

[6] We pulled all available tweets associated with notes locked as "helpful," "unhelpful," or "needs more ratings," including the date, the text of the tweet, and the user ID of the Twitter account that shared the content flagged by Note writers. Out of the 3,094 unique tweets that had an attached Community Note, 929 of these tweets written by 733 unique Twitter accounts were still available on Twitter in mid-March 2023. Out of the 32,886 unique tweets that received a Community Note with the locked status of "needs more ratings" or "unhelpful," meaning that the note was not displayed along with the tweet, 9,523 of these tweets written by 5,995 unique Twitter accounts were still available.

[7] Excluding retweeted content, we pulled a total of 138,235 additional tweets.

[8] We were able to estimate the partisan ideology for 569 Twitter users (or around 80% of the sample). We classify the remaining 164 accounts as "Unknown." This approach relies on the assumption that users follow individuals with whom they are ideologically similar and has been a standard strategy for estimating the political ideology of Twitter users in computational social science research.

[9] We were able to retrieve information about the change in followers across the weeks before and after receiving a note for 703 account-weeks.

notes that reached “helpful” status tripled, to 12.5%. The frequency with which notes reach a definitive locked status is likely to further improve due to recent alterations to the Community Note’s ranking algorithms (Twitter 2023c). Despite these changes, it will also be critical to ensure that the communities’ time and resources are allocated toward addressing Twitter’s most contentious content. We found several examples of notes attached to outdated content, including a tweet from the @Theranos account, a once-prominent healthcare company charged with fraud by the SEC in 2018.[10] The tweet was posted in 2015—six years before Community Notes’ pilot program began. Given the extremely limited resources of the Community Notes community, contributor time would likely be better spent elsewhere.

With respect to the estimated partisan distribution of users who received Community Notes attached to their tweets, we found that prior to the expansion of the program, users estimated to be more liberal received around 50% of all notes, whereas users estimated to be more conservative received 24%. After the program expanded, this partisan gap has closed from more than 25% to around 12%.[11] This is striking given that the expansion of the program coincided with Elon Musk’s takeover of Twitter, a topic frequently flagged for additional context by Community Notes contributors in the tweets of more liberal users. Removing these tweets from the dataset further shrinks the partisan gap from 12% to 8%. This suggests that expanding the contributor base to more participants has leveled the partisan balance of the Community Notes program.

Despite these more positive shifts, crowdsourced content moderation solutions such as Community Notes are limited in important ways. After examining the 929 tweets still available online with a Community Note in our dataset and manually assigning classifications to each tweet, we found that the most common reason a tweet received a “helpful” Community Note was because it contained a doctored or misleading image, video, or quote. These tweets were not strictly political, and included historical information, fabricated quotes and tweets by prominent figures, and images used out of context. However, their prominence on Twitter—and regular fixture as targets of the Community Notes program—is troubling in light of recent developments in generative AI, including language models like ChatGPT and image generators like DALL-E 2 or Midjourney. These advances have lowered the cost and difficulty of producing disinformation, and their outputs are becoming far more realistic. To effectively push back against AI-generated text, images, and even videos that spread across Twitter, the program will need to scale significantly and devise additional solutions beyond those already in play to better contend with fabricated content.

Community Notes also does not seem to impact the subsequent behavior of users. While data limitations make a program evaluation difficult, we can assess various aspects of the program, including its effect on tweet deletion rates, engagement statistics, and follower behavior, before and after receiving a Community Note.

Unlike Twitter Notices,[12] Community Notes does not force deletion or reduce the visibility of tweets containing misleading information; yet we found ample examples of tweets containing clearly fabricated content—with notes denoting as much—that remained

[10] Theranos (@theranos). Twitter post: *When you walk in for a Theranos lab test, we treat you as we’d like to be treated. As a guest, not a number.* December 9, 2015, URL: https://twitter.com/theranos/status/674717899802591232. A note that has since been removed from the post read: “This tweet was and is dishonest. Elizabeth Holmes and her company Theranos took millions of dollars from gullible investors and more from trusting consumers. It was a fraud scheme. Lawsuits resulted. [URL]. Holmes was convicted of felonies and sentenced to federal prison. [URL][URL].” Screenshot available on request.

[11] We were unable to estimate the partisan ideology based on their Twitter following choice for 164 accounts (≈22%).

[12] Notices are actions taken or context provided by Twitter’s systems and teams on tweets that violate the Twitter Rules.

online as of May 4, 2023.[13] One might expect the authors of these tweets to delete them after receiving a note clearly labeling their content as fake, but Community Notes does not appear to have this effect. While many tweets are eventually deleted, in our data, we found that the deletion rate for tweets that received a note marked as "helpful"—and as a result, displayed below the tweet in question—was no different from the deletion rate for tweets that received notes either marked as "unhelpful" or "needs more rating," and as a result were not publicly visible to users outside the Community Notes program.

In addition to a user being no more likely to delete their tweet, we found that on the day a Twitter user receives a note on one of their tweets, average likes are higher than the days before and after. This increase in engagement seems to be driven by smaller accounts. By contrast, users with over five million followers who receive a Community Note on a tweet do not experience elevated engagement. Across all users, a Community Note seems to have no impact on subsequent engagement levels. This suggests that users do not approach accounts that receive Community Notes on their tweets more skeptically in subsequent interactions.

Based on these trends it is possible that Community Notes target viral tweets from relatively small accounts, putting the content moderation program at odds with Twitter's own recommender system. Although tweets that received a content warning from Twitter have historically been deamplified, it is unclear what role Community Notes plays in mitigating tweet virality. Given that tweets that receive Community Notes are no less likely to be amplified, in some cases, they might even exacerbate this virality by drawing more attention to a misleading tweet (Twitter 2023a).

Finally, in addition to increased engagement, Twitter users who received a Community Note attached to one of their tweets also saw a boost in followers. Drawing on data from 703 account weeks, we found that the median boost in followers the week a Community Note appeared on a tweet is around 500. Although this pattern is most evident for accounts with a smaller number of followers, it does seem to persist for accounts of all follower ranges, representing an anomaly of a week in terms of follower gain for the Twitter users. Again, based on available data, it is unclear how the note factors into this increased audience and what role virality plays in the expansion of users' follower networks. If the note halted the viral spread of this content and subsequent increase in followers, an expansion of the program will allow these notes to be attached to tweets faster and thus stem their reach more rapidly.

4 The Future of Crowdsourced Content Moderation

Due to contentious debates around content moderation and cost-cutting measures across the tech sector, crowdsourced solutions are likely to expand their footprint across the information ecosystem. As the Community Notes program has expanded, it has shown some promise as a crowdsourced approach. However, there are several limitations that hinder its effectiveness and will likely remain challenges for other community-based solutions as well. Based on our findings, the impact of Community Notes—and future iterations of the content moderation mechanism across other social media websites—will need to contend with a variety of challenges in order to meet demands for online content moderation (Kozyreva et al. 2023).

[13] David Vance (@DVATW). Twitter post: *Hi @AyannaPressley - Can you please explain what this means? Thanks.*, January 17, 2023, URL: https://twitter.com/DVATW/status/1615405919944478720; Laurence Tribe (@tribelaw). Twitter post: *WTAF?! Is this an application for the Nutcase of the Year award? Or just the dumbest ad ever for IVF? https://t.co/pUHFo4M2SH*, January 29, 2023, URL: https://twitter.com/tribelaw/status/1619729915653525504.

Given the extremely small number of users who ever write a successful note, it will be critical for crowdsourced content moderation tools to identify ways to responsibly and rapidly scale their reach, whether by limiting the time frame during which users can post contributions or requiring qualified users to opt out of the program rather than opt in. Beyond issues of volume, crowdsourced content moderation faces a daunting challenge of seeking out consensus across divergent perspectives in an increasingly polarized information ecosystem. With presidential elections fast approaching in 2024, election-related disinformation will likely remain a persistent threat. Given the difficulty of reaching a consensus already, disagreements across partisan lines will likely hinder any ability for crowdsourced solutions that hinge on consensus to counteract the spread of this content unless they scale rapidly or deploy other systems to override consensus-generating mechanisms on particularly partisan topics. Further experimentation with decreasing the requirements for agreement, while guarding against the partisan targeting flagged by prior research, or creating a tiered system of moderation like Wikipedia with some “super users” who are able to overrule particularly egregious notes could help sidestep consensus for claims that are pressing but may never reach agreement across “divergent viewpoints.”

Additional challenges also loom. Given the prominence of doctored and decontextualized images and videos across social media platforms such as Twitter, text generators like ChatGPT and image generators like Dall-E or Midjourney will create a world where fake images, text, and videos spread even more rapidly and are far more difficult for untrained users to detect on their own. We have already seen this occur, after an AI-generated image of an explosion near the Pentagon caused stock markets to plunge (Marcelo 2023). Content moderation tools—including crowdsourced solutions—must be equipped to deal with this effectively. Platforms could consider removal of doctored or false images, quotes, and videos, particularly if they are used without explicitly stating that they are generated or doctored. But these solutions will likely not remedy the challenge. Without adequate digital literacy and education efforts, experts and trained professionals will be far better positioned to detect this content than the average social media user. If and how “wisdom of the crowd” solutions effectively contend with these urgent challenges will shape their efficacy and success as content moderation practices moving forward.

References

Allen, Jennifer, Antonio A. Arechar, Gordon Pennycook, and David G. Rand. 2021. “Scaling Up Fact-Checking Using the Wisdom of Crowds.” *Science Advances* 7 (36): eabf4393. https://doi.org/10.1126/sciadv.abf4393.

Allen, Jennifer, Cameron Martel, and David G. Rand. 2022. “Birds of a Feather Don’t Fact-Check Each Other: Partisanship and the Evaluation of News in Twitter’s Birdwatch Crowdsourced Fact-Checking Program.” In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems,* 1–19. https://doi.org/10.1145/3491102.3502040.

Barberá, Pablo. 2015. “Birds of the Same Feather Tweet Together: Bayesian Ideal Point Estimation Using Twitter Data.” *Political Analysis* 23 (1): 76–91. https://doi.org/10.1093/pan/mpu011.

Borak, Masha. 2022. “The Hunt for Wikipedia’s Disinformation Moles.” *Wired,* https://www.wired.com/story/wikipedia-state-sponsored-disinformation/.

Borwankar, Sameer, Jinyang Zheng, and Karthik Natarajan Kannan. 2022. “Democratization of Misinformation Monitoring: The Impact of Twitter’s Birdwatch Program.” *Available at SSRN 4236756,* https://doi.org/10.2139/ssrn.4236756.

Dang, Sheila. 2022. “Exclusive: Twitter Is Losing Its Most Active Users, Internal Documents Show.” *Reuters,* https://www.reuters.com/technology/exclusive-where-did-tweeters-go-twitter-is-losing-its-most-active-users-internal-2022-10-25/.

Delcker, Janosch. 2022. “Twitter’s Sacking of Content Moderators Raises Concerns.” *Deutsche Welle,* https://www.dw.com/en/twitters-sacking-of-content-moderators-will-backfire-experts-warn/a-63778330.

Faife, Corin. 2022. “COVID Misinfo is the Biggest Challenge for Twitter’s Birdwatch Program, Data Shows.” *The Verge* (October 10, 2022). https://www.theverge.com/2022/10/10/23393021/twitter-birdwatch-covid-misinformation-data-analysis-misinformation-fact-check.

Gallo, Jason A., and Clare Y. Cho. 2021. “Social Media: Misinformation and Content Moderation Issues for Congress.” *Congressional Research Service Report* 46662.

Godel, William, Zeve Sanderson, Kevin Aslett, Jonathan Nagler, Richard Bonneau, Nathaniel Persily, and Joshua A. Tucker. 2021. “Moderating with the Mob: Evaluating the Efficacy of Real-Time Crowdsourced Fact-Checking.” *Journal of Online Trust and Safety* 1 (1). https://doi.org/10.54501/jots.v1i1.15.

Goldberg, David A. M., Claudia Lo, and Margeigh Novotny. 2020. “Content Moderation and Anti-Vandalism: Wikipedia’s Use of AI.” *Wikimedia Design Blog,* https://www.design.wikimedia.org/blog/2020/07/30/content-moderation-anti-vandalism-wikipedia.html.

Google. 2023. “About the YouTube Priority Flagger program.” *Google,* https://support.google.com/youtube/answer/7554338?sjid=10340825454164077783-NA.

Harwell, Drew. 2022. “A Fake Tweet Sparked Panic at Eli Lilly and May Have Cost Twitter Millions.” *The Washington Post,* https://www.washingtonpost.com/technology/2022/11/14/twitter-fake-eli-lilly/.

Howe, Jeff. 2006. “The Rise of Crowdsourcing.” *Wired,* https://www.wired.com/2006/06/crowds/.

Klonick, Kate. 2017. “The New Governors: The People, Rules, and Processes Governing Online Speech.” *Harvard Law Review* 131:1598.

Kozyreva, Anastasia, Stefan M. Herzog, Stephan Lewandowsky, Ralph Hertwig, Philipp Lorenz-Spreen, Mark Leiser, and Jason Reifler. 2023. “Resolving Content Moderation Dilemmas Between Free Speech and Harmful Misinformation.” *Proceedings of the National Academy of Sciences* 120 (7): e2210666120. https://doi.org/10.1073/pnas.2210666120.

Mahadevan, Alex. 2021. “Analysis: Twitter’s Crowdsourced Fact-Checking Experiment Reveals Problems.” *Poynter Institute for Media Studies,* https://www.poynter.org/fact-checking/2021/analysis-twitters-crowdsourced-fact-checking-experiment-reveals-problems/.

Marcelo, Phillip. 2023. “FACT FOCUS: Fake Image of Pentagon Explosion Briefly Sends Jitters Through Stock Market.” *Associated Press,* https://apnews.com/article/pentagon-explosion-misinformation-stock-market-ai-96f534c790872fde67012ee81b5ed6a4.

Peoples, Steve, Anthony Izaguirre, and Adriana Gomez Licon. 2023. “DeSantis Launches GOP Presidential Campaign in Twitter Announcement Plagued by Glitches.” *Associated Press,* https://apnews.com/article/ron-desantis-2024-presidential-election-live-updates-0495d567326db1e760179d01f1f7c45e.

Renfro, Kim. 2022. “For Whom the Troll Trolls: A Day in the Life of a Reddit Moderator.” *Business Insider,* https://www.businessinsider.com/what-is-a-reddit-moderator-2016-1.

Singh, Spandana. 2019. “Everything Moderation: An Analysis of How Internet Platforms are Using Artificial Intelligence to Moderate User-Generated Content.” *New America,* https://www.newamerica.org/oti/reports/everything-moderation-analysis-how-internet-platforms-are-using-artificial-intelligence-moderate-user-generated-content/.

Stokel-Walker, Chris. 2022. “Reddit Moderators do $3.4 Million Worth of Unpaid Work Each Year.” *New Scientist,* https://www.newscientist.com/article/2325828-reddit-moderators-do-3-4-million-worth-of-unpaid-work-each-year.

Stuart, S.C. 2021. “Wikipedia: The Most Reliable Source on the Internet?” *PCMAG,* https://www.pcmag.com/news/wikipedia-the-most-reliable-source-on-the-internet.

Surowiecki, James. 2004. *The Wisdom of Crowds: Why the Many Are Smarter Than the Few and How Collective Wisdom Shapes Business, Economies, Societies and Nations.* Doubleday; Anchor. https://psycnet.apa.org/record/2004-20179-000.

Twitter. 2022. “Helpful Birdwatch Notes Are Now Visible to Everyone on Twitter in the US.” *Twitter* (October). https://blog.twitter.com/en_us/topics/product/2022/helpful-birdwatch-notes-now-visible-everyone-twitter-us.

———. 2023a. “About Community Notes on Twitter.” *Twitter,* https://help.twitter.com/en/using-twitter/community-notes.

———. 2023b. “Diversity of Perspectives.” *Twitter,* https://communitynotes.twitter.com/guide/en/contributing/diversity-of-perspectives.

———. 2023c. “Note Ranking Algorithm.” *Twitter,* https://communitynotes.twitter.com/guide/en/under-the-hood/ranking-notes#modeling-uncertainty.

———. 2023d. “Notes shown on Twitter.” *Twitter,* https://communitynotes.twitter.com/guide/en/contributing/notes-on-twitter.

Wikipedia. n.d. “Wikipedia: Administrators.” *Wikipedia,* accessed August 16, 2023. https://en.wikipedia.org/wiki/Wikipedia:Administrators.

Authors

Valerie Wirtschafter is a senior data analyst at the Brookings Institution. She received her PhD in political science from the University of California, Los Angeles in 2021.

(valerie.wirtschafter@gmail.com)

Sharanya Majumder is a project assistant at the Brookings Insitution.

Acknowledgements

We thank Chris Meserole, Jessica Brandt, and Derek Belle for their comments and feedback on various drafts.

Keywords

Crowdsourcing; fact-checking; content moderation; misinformation.