Backlash or Bullying? Online Harassment, Social Sanction, and the Challenge of COVID-19 Misinformation

Timothy J. Foley and Melda A. Gurakar

Abstract. Online platforms continue to grapple with the spread of false information about the COVID-19 pandemic, especially about the safety and effectiveness of the COVID-19 vaccine. Some users who disseminate vaccine misinformation report that they were bullied by other users in response to their anti-vaccine messages. When they arise, these reports pit a platform's prerogative to reduce the spread of misinformation against its obligation to protect users from online harassment. To resolve this tension, we present a framework that evaluates user interactions based on three criteria: intensity, specificity, and persistence. This approach can help content moderators determine when other users' criticism of anti-vaccine messages constitutes harassment. After exploring the framework and its theoretical underpinnings, we report the results of an experimental survey (n=21) that compares moderation decisions made using this new policy framework for our social media platform, Patio, to those made based on our existing community guidelines. We find that the framework yields a statistically significant improvement in the overall accuracy and precision of moderation decisions involving the potential harassment of users spreading COVID-19 vaccination misinformation. We conclude by considering the limitations of our analysis and avenues for further research.

1 Introduction

On August 2, 2021, the United States Court of Appeals for the Seventh Circuit upheld Indiana University's COVID-19 vaccine mandate for its nearly 90,000 undergraduate students (*Klaassen v. Indiana University* 2021). This ruling cleared the way for schools nationwide to require their students and staff to be vaccinated against the virus. As of August 2021, over 800 colleges and universities had mandated the COVID-19 vaccine for residential students (Camera 2021). Days after this court decision, an incoming student at a large public university far from Chicago rejoiced that their campus would also require vaccination. The student posted on a campus-wide group chat on Patio, an emerging social network, that it "took [their school's administrators] long enough." This comment

10.54501/jots.v1i2.31

ignited a firestorm of online discussion; some students supported the decision, while others opposed any institutional mandate to receive the vaccine.

While platforms work to minimize the spread of misinformation, they also prohibit bullying and harassment online. On some platforms, these goals can come into conflict when users who spread misinformation about the COVID-19 vaccine report comments from supporters of the vaccine that they argue constitute bullying. However, the case was complicated because what vaccine-hesitant users viewed as bullying could also be considered social sanction in response to spreading misinformation. On the one hand, users should be dissuaded from sharing misinformation, and criticism from other users is an effective way to do that. On the other hand, users do not have *carte blanche* to say whatever they wish to other users whose views they do not agree with. How should platforms strike the delicate balance between a prohibition on bullying and an obligation to limit the spread of vaccine misinformation?

Patio is a message-based social network platform on which users directly converse with one another in chats of varying sizes. Its early users have included students on college campuses using college-wide chats. Its community guidelines prohibit bullying and harassment on the platform (Patio 2021). The company also considers a responsibility against misinformation. Using the company's internal data (anonymized for privacy), we present a new policy approach that provides actionable steps for content moderators to balance the competing prerogatives endemic to platform governance.

This analysis is based on actual content moderation decisions made by Patio's Trust and Safety team. College students were particularly impacted by COVID-19, as they were forced to evacuate their dorms and shift their education online early in the pandemic (Hess 2020). Many students took an interest in their university's policies for managing the pandemic's impact on the classroom. In the lead-up to the Fall 2021 semester, frequent discussions of such policies sprang up on campuses across the United States. Some of these discussions turned heated and involved a mix of misinformation and harassment. Sensing a novel policy challenge, Patio's Trust and Safety team examined several cases to develop the policy approach we describe here. It is important to note that the platform employs a "commercial moderation" model, in which employees specializing in trust and safety both develop policies and make content moderation decisions (Roberts 2016). All content moderators are full-time Patio employees. Decisions about whether to remove a message may depend on the context in which it was sent, which moderators analyze by considering the messages preceding and following the reported content. We sought to create a framework that allows content moderators to resolve the tension between their dual commitments to protect users from misinformation as well as bullying and harassment. This framework, created by and for industry, represents a practical solution to an underexplored problem in content moderation.

This study contributes to the scholarly literature on online bullying and harassment. Long a subject of academic attention, research on cyberbullying exploded in the 2010s after the mass adoption of social media platforms among young people. A meta-analysis conducted by Kowalski et al. surveyed 131 papers on cyberbullying that studied the psychological effects, educational outcomes, and technological systems involved in online bullying and harassment (Kowalski et al. 2014). Because social media became a major vector for online bullying, industry and academic attention shifted to analyzing platforms' responsibility to crack down on online harassment. Chan et al. examined research on cyberbullying as it relates to social media networks and found that prevention, detection, and enforcement via platform policies were major foci of academic and industry research (Chan, Cheung, and Lee 2021). Milosevic's 2016 survey of company policies against bullying raised concerns about whether platforms were equipped to clamp down on abuses perpetrated on their platforms (Milosevic 2016). Blackwell et al. evaluated

whether users believe online harassment can be justified based on a person's past actions, but did not propose a specific platform policy response (Blackwell et al. 2018). Our study builds upon this body of literature by analyzing the complex balance that companies must strike to effectively limit misinformation and abuse on their platform.

This study proposes and tests a content moderation framework that maintains the company's responsibility to limit the spread of misinformation *and* upholds the platform's prohibition of bullying and harassment. We begin by arguing that social sanction, defined as the "expression of disapproval with a particular kind of conduct," can help limit the spread of misinformation online (Nelissen and Mulder 2013). We proceed in two parts. In Part I, we delineate a tripartite framework for distinguishing social sanction from harmful bullying and harassment. This framework contains three axes: intensity, specificity, and persistence. In Part II, we test this framework on real discussions among Patio's users, and find that it provides clear and consistent guidance to content moderators.

Our experiment compared our policy framework to our pre-existing community guidelines. We selected two sets of messages from actual group chats between college students on Patio. We recruited volunteers to serve as content moderators and randomly divided them into two groups. One group was asked to decide whether any user messages from these group chats should be removed from the platform after reading Patio's community guidelines, while the other group was asked to decide whether any content (from the same group chats) should be removed after reading our policy explanation. We then compared the number of violations reported by both groups. Overall, the policy explanation framework performed better than the community guidelines. Prior to the study, Patio's Trust and Safety team determined that only one message from the sample analyzed by the volunteers violated the platform's policies; 90% of respondents in the policy explanation group recommended removing the offending message, compared to 77% of the community guidelines group. The standard deviation of the policy explanation group was also lower than that of the community guidelines group, which indicates more consistent results.

2 Distinguishing Social Sanction from Bullying

Internet platforms like Patio create norms for their users through their community guidelines. Patio's Trust and Safety team modeled its community guidelines after the existing best practices in the field; they prohibit users from sending spam, imitating other people, and harassing other users (Patio 2021). These guidelines are dynamic and are periodically updated in response to emerging content moderation challenges. The company also prohibits users from spreading harmful misinformation related to the COVID-19 vaccine. If these rules are violated, the company's Trust and Safety team may take action to remove the offending content or users.

Decisions to act against a user who breaks the platform's rules are made based on the context of the reported situation, but platforms should ensure that such assessments are made consistently. Platforms should generally strive to treat each user equally in content moderation decisions. Arbitrarily favoring one user over another or denying a user accused of violating community guidelines the appropriate consideration amounts to arbitrary discrimination, which infringes on a user's right to access information, freely express their ideas, and participate in the online ecosystem (Dias Oliva 2020).

Whether a user is spreading public health misinformation is relevant in content moderator decisions. Misinformation can cause real harm if it dissuades users from getting vaccinated. This is true among Patio's users, many of whom attend colleges and universities. In a recent survey, almost half (47.5%) of those enrolled in higher education

reported being hesitant to obtain the COVID-19 vaccine (Sharma, Davis, and Wilkerson 2021). Misinformation can amplify and intensify this hesitancy, and risks reducing the vaccination rate among a critical demographic—which could further prolong this deadly pandemic.

Preserving community norms, such as those against misinformation, can require sanctioning those who violate them (Blake and Davis 1964). While there is a consensus that norms require some type of sanction to effectively guide actions both on and off the internet, simply violating a norm does not give the collective—a society, corporation, or group of individuals—free reign to do whatever they wish to the violator. Proportionality has become a significant factor in the philosophy of punishment, especially in criminal matters (Hirsch 1992). Platform users can socially sanction each other by expressing disapproval of their decision to spread false information. Our platform seeks to ensure that (1) there is a rough correspondence between the action that violated communal expectations and our moderation response and that (2) group members do not have free reign to take revenge on the offending subject.

There is a distinction, however unclear to both platforms and users, between permitted social sanctions levied by other users on the one hand, and unjustifiable bullying on the other hand. We propose a framework to determine whether a user's actions constitute harassment. The framework assesses three factors: intensity, persistence, and specificity. Intensity denotes the severity of the comment, persistence captures the frequency of comments directed at the target, and specificity measures the size of the aggressor's intended audience and whether the targeted person or group is easily identifiable. These factors are balanced against each other to determine whether content should remain on Patio's platform. This framework is visualized on a three-dimensional axis: the grey pyramid Figure 1 on the next page displays the estimated range of permitted content.

Content moderation decisions require considering all three criteria. For example, repeated low-intensity, medium-specificity comments may justify removal, just as a single comment of medium intensity and high persistence encourages moderation action. Yet a comment with low specificity and low intensity would need to be reposted more often than one with high specificity and low persistence to justify platform action. We suggest that direct threats against a singular person or small group are more harmful than a pattern of aggressive comments directed at a general group for two reasons. First, police departments, especially on college campuses, have long held that the highest-level threats are "direct, specific, and plausible" (UALR, n.d.). This implies that naming a target, as opposed to insulting or condemning a large group, is considered a greater threat. The more specific and detailed the target of a threat is, the more danger they are in.[1] In other words, it is easier for an individual to harm another individual than to take on an entire group at once. Second, condemning or harassing a particular person or specific group (e.g. a club) isolates them from their peers and marks them for scorn. This isolating effect can expose them to second-order harms, as is common in cases of doxxing which, by its nature, targets a specific individual or group (Anderson and Wood 2021). This explains why the threshold for removal is higher on the persistence axis, lower on the intensity axis, and lowest on the specificity axis.

1. It is important to note that threats against general groups and other forms of hate speech, especially communities protected on the basis of race, ethnicity, or gender, can (and should) be taken seriously. Our notion of specificity only suggests that it is an important factor to consider in cases of bullying and harassment against individuals or groups. We do not seek to minimize in any way the challenge that hate speech poses to content moderators.

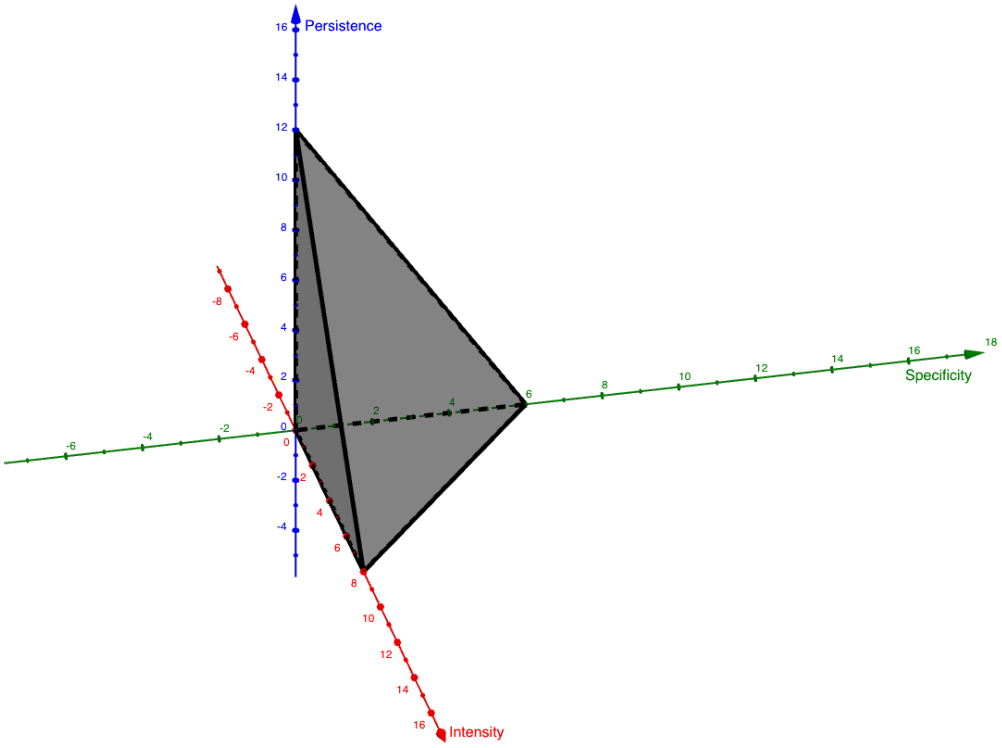

Figure 1: Intensity is shown on the x-axis, specificity on the y-axis, and persistence on the z-axis. The grey pyramid denotes protected comments to balance the three factors against one another.

3 Applications to Online Discussion of COVID-19 Vaccinations

We tested our framework's ability to (1) guide moderators to make accurate decisions and (2) promote consistent moderation practices. We consider a result to be *accurate* when the volunteer moderators removed the same message as Patio's Trust and Safety team, and *consistent* when the volunteer moderators all detected the same number of violations.

3.1 Methodology

We began by analyzing Patio's internal dashboard for vaccine-related discussions that involved one user reporting another for bullying or harassment. We surveyed user messages from six large, public universities sent in public groups between July, 31 and August 9, 2021. One university is in the Southwest, two are in the Southeast, and three are in the Midwest. During each of these conversations, a Patio user who expressed vaccine hesitancy or shared vaccine-related misinformation reported another user's message for harassment. From these six university chats, we selected two cases to use in our experiment to test the effectiveness of our framework. We selected these cases because they contained misinformation regarding the COVID-19 vaccine and led one user to report another for bullying and harassment. Each case consisted of approximately 20 messages (10 before and 10 after the reported message). To protect user privacy, all personally identifiable information—including participant and school names, profile pictures, time stamps, and the number of likes the messages received—were redacted before they were sent to the research respondents.

After considering both the content of the individual messages and the overall context of the conversation, our professional moderators assessed that there were no reported messages of bullying and harassment that violated the community guidelines in Case 1 and one message that violated the new framework guidelines in Case 2. This was our baseline for accuracy. Although moderation decisions are rarely clear cut, and often depend to some extent on the predispositions of the moderator, we believe some comments are clearer violations than others. By identifying the number of comments that we believe violate the platform's policies, we can then analyze which group comes closer to this number to determine which approach yields a more accurate result relative to the "true" number of violations contained in the sample.

In the second case, a user wrote in a school-wide chat that she refused to get the vaccine on the basis that it was unhealthy. In response, another user stated that physical violence was appropriate against the unvaccinated because they put the broader community at risk. This user's message was reported to Patio for bullying; it was deemed to violate the community guidelines and was removed. The second case occurred when a user claimed that the vaccine was a conspiracy for the government to control the population. Other users responded that this view was dangerous to the public health. These users were reported for bullying, but Patio concluded that no moderator response for bullying was necessary for Case 2.

After the cases were selected and compiled, we solicited a group of respondents to act as content moderators (n = 21) from among the authors' professional and social networks. Respondents were first asked if they would be willing to complete a 20-minute survey testing the effectiveness of Patio's content moderation policy. Because the respondents were not randomly selected from the population, we assigned each subject to a group at random to preserve statistical integrity. The Community Guidelines group (the statistical control group) was given Patio's publicly available community guidelines (see Appendix A). The Policy Explanation group received a copy of the policy framework described above

(see Appendix B).

Respondents were then sent a link to an online survey (reproduced in Appendix C). Both groups were given the cases discussed above and asked to assess whether any comments violated Patio's community guidelines, how many violations occurred, and to rate their confidence on a scale from 1 to 5. While the survey was anonymous, we did collect basic demographic information such as age, gender, vaccination status, and occupation to control for potential confounding variables that may create differences between groups. The survey responses were automatically recorded, downloaded, and then imported into R for analysis.

3.2 Results

The results of the tests are presented in the following tables. We began by analyzing the data for outliers, identifying two statistical anomalies—one in the control data and one in the response data.[2] After removing these two outliers, our total sample size was 21 moderators: 12 in the Community Guidelines group and 9 in the Policy Explanation group (see Table 1).

Table 1: Distribution of respondents

	Community Guidelines			Policy Explanation		
Statistic	**Case 1**	**Case 2**	**Total**	**Case 1**	**Case 2**	**Total**
n	12	12	12	9	9	9

Respondents in the Community Guidelines group reported more violations than those in the Policy Explanation group. Additionally, the standard deviation was higher in the former than the latter, implying more varied and inconsistent results.

Table 2: Summary statistics

	Community Guidelines			Policy Explanation		
Statistic	**Case 1**	**Case 2**	**Total**	**Case 1**	**Case 2**	**Total**
SD	1.75	0.9	2.02	0	0.667	0.667
Mean	1.692	1.083	2.417	0	0.778	0.778
Min	0	0	0	0	0	0
1st Quartile	0	0.75	0.75	0	0	0
Median	1	1	2	0	1	1
3rd Quartile	3	1.25	4	0	1	1
Max	6	3	6	0	2	2

The first test we conducted was Levene's Test for the homogeneity of two variances. This helped determine whether the policy explanation made moderation decisions more consistent in a statistically significant manner (Snedecor and Cochran 1989). We employed the following hypotheses:

2. Both outliers exceeded three times the interquartile range of subject-assessed violations at the higher bound. These outliers can be understood as the product of abnormal data collection conditions, as our survey was not optimized for mobile devices, yet one subject responded on their cell phone. The other reported rushing through the survey, which implies the data is not reliable. We then analyzed the data for normality, concluding that the data of subject-assessed violations was not normally distributed according to the Shapiro-Wilk normality test. This required us to utilize nonparametric tests that do not rely on an assumption of normality.

$$H_0 : \sigma_c = \sigma_p$$
$$H_A : \sigma_c \neq \sigma_p$$
$$\alpha = 0.05$$

Where *c* denotes the community guidelines group and *p* denotes the policy explanation group and * denotes statistical significance when α=0.05. The results, reported in Table 3, indicated that the policy explanation exerts a statistically significant effect on moderation consistency for Case 1 and overall, but not for Case 2. Because the p-value is less than the alpha, we reject the null hypothesis and accept the alternative that the variance in the results for the Community Guidelines group is greater than that for the Policy Explanation group for Case 1 and overall, but not for Case 2.

Table 3: Results of the Levene's test

Statistic	**Case 1**	**Case 2**	**Total**
p-value	0.0078*	0.6133	0.0186*
F value	8.7576	0.264	6.622
d.f.	20	19	19

Levene's test, while typically used to test the assumption of equal variances, can be interpreted in this context as indicating that the groups have different variances. The unequal variance can be interpreted to indicate that the Policy Explanation group returned more consistent results than the Community Guidelines group. This conclusion is further supported when we compare the standard deviation of the Policy Explanation group to the Community Guidelines group. In Case 1 and overall, the standard deviation was higher in the Community Guidelines group than the Policy Explanation group. This analysis of the standard deviation shows us the *direction* of that difference, and demonstrates that the Policy Explanation group is more consistent than the Community Guidelines group.

We also tested for accuracy by conducting a Mann-Whitney U test for two independent samples. Recall that accuracy in this context refers to the subject's likelihood of choosing the message deemed by Patio's team to have violated the platform's community guidelines. This nonparametric alternative to the Student's *t*-test assesses whether the distribution of one sample significantly diverges from the distribution of the other sample. Because there were ties in the data, we programmed R to engage in continuity correction. This test was performed according to the following hypotheses:

$$H_0 : P(x_c > x_p) = \frac{1}{2}$$
$$H_A : P(x_c > x_p) > \frac{1}{2}$$
$$\alpha = 0.05$$

Our alternative hypothesis is that the Community Guidelines group shifted to the right of the Policy Explanation group, reflecting a higher rate of violations assessed by people who only had access to our community guidelines. The results of the test indicate a statistically significant difference between the distribution of both groups. Thus, in Case 1 and overall, we reject the null in favor of the alternative but not in Case 2 (see Table 4 on the next page).

Table 4: Results of the Mann-Whitney U test

Statistic	Case 1	Case 2	Total
p-value	0.0013*	0.2416	0.0289*
W	99	63.5	80.5

The Mann-Whitney U test proves that the distribution of detected violations in Case 1 and overall is different between the two groups. This finding suggests that our new policy explanation did cause moderators to act differently than those who consulted the community guidelines. However, this test alone cannot prove the accuracy of the policy explanation framework because the population did not have equal variances (McKnight and Najab 2010). To determine whether this difference resulted in more or less accurate decisions, we examine the difference in the average number of violations reported between the two groups. Because the Policy Explanation group detected a lower average number of violations than the Community Guidelines group, we can conclude that the former is more accurate than the latter.

To further support our hypothesis, we created a density distribution for Case 1 (Figure 2) and total violations (Figure 3 on the next page). We did not include a density distribution for Case 2, as both groups reported the same number of violations. In Case 1, the Policy Explanation group delivered more consistent results, as evidenced by the narrow density distribution. It also reported an average number of violations much closer to the accurate number, determined prior to the experiment by Patio's Trust and Safety team.

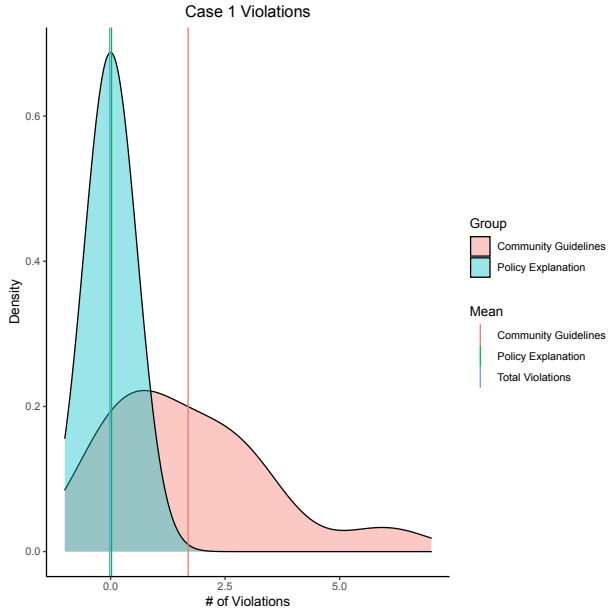

Figure 2: Density distribution for Case 1

Overall, the Policy Explanation group was more consistent and more accurate than the Community Guidelines group. The average number of violations detected by the Policy Explanation group (green line) was significantly closer to the predetermined parameter (blue line) than those reported by the Community Guidelines group (pink line).

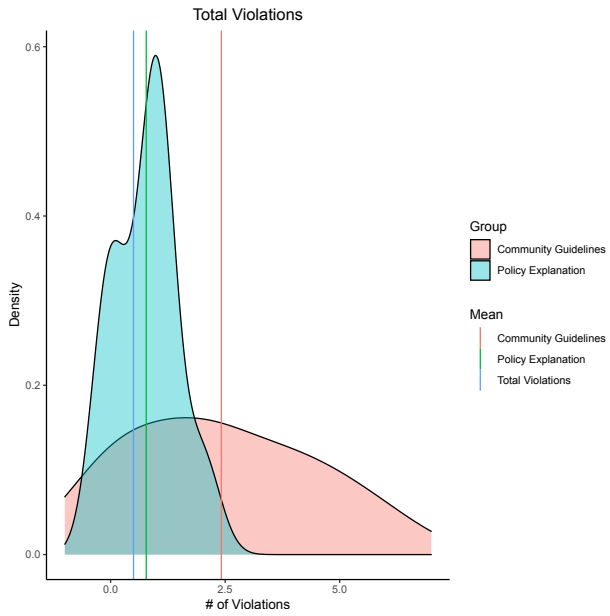

Figure 3: Density distribution (overall)

In each figure, we marked the average number of messages that respondents thought violated our community guidelines, as well as the predetermined number of messages that were actual violations with vertical lines. In both Case 1 and overall, the Policy Explanation group is closer to the total number of violations than the Community Guidelines group. Taken together with the Mann-Whitney U test, this implies that the former yielded more accurate results in Case 1 and overall.

3.3 Discussion

Our results indicate that our policy framework led to more accurate and precise (consistent) decisions in both Case 1 and overall. It did not yield the same effects for Case 2, because the message that constituted a clear violation of community guidelines occurred in Case 2 and was selected for removal by both groups at high rates. This suggests that obvious policy violations can be detected reasonably accurately regardless of whether a moderator uses the community guidelines or the more detailed policy explanation. The benefits of the policy explanation stem from its predisposition towards under-moderation. Because the respondents using the policy framework reported more comments as violative than their community guideline counterparts, the policy framework leaves more messages on the platform than a broad application of the community guidelines.

4 Conclusion

This article presented and tested a framework designed to resolve the conflict between accuracy and precision in social media platforms' moderation decisions. Our experiment indicates that using the policy explanation yields a statistically significant increase in both the accuracy and precision of content moderation decisions compared to just the community guidelines.

Future research could consider four possible methodological changes. First, follow-on studies could use a larger sample of respondents. Second, we note that our sample was

not randomly selected from the population. While we do not believe random selection is necessary in this instance as long as random *assignment* is preserved, it could provide an interesting area for future research. Given that content moderation is often considered "unskilled" labor, understanding how the general population views this work could yield new insights into platform policy (Newitz 2020). Third, testing this framework on additional cases could yield more robust results that could help further verify the validity of our conclusions. Finally, our framework was tailored to message-based platforms, which feature direct user communication without a central post. It is less clear whether it would be as effective on post-based platforms like Facebook or Twitter, although we expect it could be as effective when applied to the comment section or reply threads of those platforms.

Further research is needed on a wide array of questions in platform governance. For example, research continues on the role of social sanctions in quelling the spread of misinformation. While public condemnation certainly has an effect, the depth and significance of this effect is not yet well understood. Additionally, companies rarely submit their policy enforcement frameworks for peer review, and there is no clear best practice in the industry to test policy efficacy. Establishing methods to develop, test, and implement consistent moderation policy should be a prominent area of academic and industry cooperation.

Platforms continue to face a confluence of challenges resulting from the COVID-19 pandemic, including how to handle users who spread misinformation or express vaccine hesitancy. Platforms have a potentially outsized influence over whether some members of the public choose to get vaccinated. All platforms that traffic in public information have an obligation to acknowledge this supporting role, and must be prepared to resolve the tensions and conflicts within their own sets of rules to protect the networked society they inform.

References

Anderson, Briony, and Mark A Wood. 2021. "Harm Imbrication and Virtualised Violence: Reconceptualising the Harms of Doxxing." *International Journal for Crime, Justice and Social Democracy* 10 (4).

Blackwell, Lindsay, Tianying Chen, Sarita Schoenebeck, and Cliff Lampe. 2018. "When online harassment is perceived as justified." In *Proceedings of the Twelfth International AAAI Conference on Web and Social Media (ICWSM 2018),* vol. 12. Association for the Advancement of Artificial Intelligence.

Blake, J, and K Davis. 1964. "Norms, values, and sanctions," 456–84.

Camera, Lauren. 2021. "School Vaccine Mandates: Here They Come | National News." *US News & World Report* (August). Accessed January 19, 2022. https://www.usnews.com/news/national-news/articles/2021-08-31/school-vaccine-mandates-here-they-come.

Chan, Tommy K.H., Christy M.K. Cheung, and Zach W.Y. Lee. 2021. "Cyberbullying on social networking sites: A literature review and future research directions." *Information & Management* 58, no. 2 (March): 103411. https://doi.org/10.1016/j.im.2020.103411.

Dias Oliva, Thiago. 2020. "Content Moderation Technologies: Applying Human Rights Standards to Protect Freedom of Expression." *Human Rights Law Review* 20, no. 4 (December): 607–40. Accessed October 18, 2021. https://doi.org/10.1093/hrlr/ngaa032. https://doi.org/10.1093/hrlr/ngaa032.

Hess, Abigail Johnson. 2020. "How coronavirus dramatically changed college for over 14 million students." Section: Make It - Work, *CNBC* (March). Accessed October 8, 2021. https://www.cnbc.com/2020/03/26/how-coronavirus-changed-college-for-over-14-million-students.html.

Hirsch, Andrew von. 1992. "Proportionality in the Philosophy of Punishment." *Crime and Justice* 16:55–98. http://www.jstor.org/stable/1147561.

Kowalski, Robin, Gary Giumetti, Amber Schroeder, and Micah Lattanner. 2014. "Bullying in the Digital Age: A Critical Review and Meta-Analysis of Cyberbullying Research Among Youth." *Psychological bulletin* 140 (February). https://doi.org/10.1037/a0035618.

McKnight, Patrick E., and Julius Najab. 2010. "Mann-Whitney U Test." In *The Corsini Encyclopedia of Psychology,* 1–1. American Cancer Society. Accessed October 29, 2021. https://doi.org/10.1002/9780470479216.corpsy0524. http://onlinelibrary.wiley.com/doi/abs/10.1002/9780470479216.corpsy0524.

Milosevic, Tijana. 2016. "Social media companies' cyberbullying policies." *International Journal of Communication* 15 (January): 647–67.

Nelissen, Rob M. A., and Laetitia B. Mulder. 2013. "What makes a sanction "stick"? The effects of financial and social sanctions on norm compliance." *Social Influence* 8, no. 1 (January): 70–80. Accessed October 13, 2021. https://doi.org/10.1080/15534510.2012.729493. http://www.tandfonline.com/doi/abs/10.1080/15534510.2012.729493.

Newitz, Annalee. 2020. "We Forgot About the Most Important Job on the Internet." *The New York Times* (March). Accessed November 12, 2021. https://www.nytimes.com/2020/03/13/opinion/sunday/online-comment-moderation.html.

Patio. 2021. *Community Guidelines,* July. Accessed September 23, 2021. https://www.notion.so.

Roberts, Sarah T. 2016. "Commercial content moderation: Digital laborers' dirty work." In *The Intersectional Internet: Race, Sex, Class and Culture Online,* edited by S.U. Noble and B. Tynes. Peter Lang Publishing, January.

Sharma, Manoj, Robert E. Davis, and Amanda H. Wilkerson. 2021. "COVID-19 Vaccine Acceptance among College Students: A Theory-Based Analysis." *International Journal of Environmental Research and Public Health* 18, no. 9 (April): 4617. Accessed October 18, 2021. https://doi.org/10.3390/ijerph18094617. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8123652/.

Snedecor, George W, and William G Cochran. 1989. "Statistical Methods, eight edition." *Iowa state University press, Ames, Iowa* 1191.

University of Arkansas, Little Rock. n.d. "Levels of Risk," https://ualr.edu/safety/home/emergency-management-plan/threat-assessment-team/levels-of-risk/.

Authors

Timothy J. Foley is a Policy Associate at Patio. He is also a Research Assistant at the Harvard Kennedy School's Program on Science, Technology, and Society.

Melda A. Gurakar is a technologist specializing in the intersection of digital rights and emerging technologies. She is the Head of Policy for Trust and Safety at Patio, an early-stage consumer social company.

Acknowledgements

We thank the Patio team and all of our survey participants for their time and insights generated through this research.

Data Availability Statement

Not applicable.

Funding Statement

Not applicable.

Ethical Standards

Because the study was conducted under the auspices of a company, it did not necessitate IRB approval as the company does not maintain such an institution. Care was taken to ensure that all research was conducted according to the highest ethical standards with a particular concern for maintaining user privacy.

Keywords

Content moderation; COVID-19; Bullying and harassment; Vaccine misinformation

Appendices

Appendix A: Patio's Community Guidelines

At Patio, we build social tools that are useful for the way you live, communicate, and collaborate. A big part of that mission is making sure you are safe while using Patio, which is why we're excited to share our community guidelines. We believe it's all of our responsibilities to keep our Patio healthy and safe, and that these guidelines can help us do that. By using Patio, you're agreeing to follow these community guidelines:

1. You must use your real name on Patio. This makes sure no one is being impersonated, and ensures everyone in the Patio is who they say they are.
2. You may not engage in abuse, bullying, or harassment of any person or groups of people. That behavior is not cool, and has no place on Patio.
3. You may not discriminate against, engage in hateful conduct directed at, or threaten violence or harm against any person or groups of people.
4. You may not post images or messages of the following types: threats, personal information of others without consent, nudity or graphic sexual images, intention to cause physical or emotional harm, promotion of self-harm or violence.
5. You may not spam or raid a Patio.
6. You may not use Patio for illegal purposes.

We take these guidelines seriously, and will not tolerate behavior that breaks these rules. If you see behavior that breaks these rules, please report it to our team via our in-app reporting buttons. We may take steps including issuing a warning, removing the content, removing and banning the accounts responsible, or other necessary actions.

Appendix B: Policy Explanation

Subsection B.1: Policy

According to Patio's community guidelines (effective July 1, 2021), users may not "engage in abuse, bullying, or harassment of any person or groups of people." Whether moderator action is necessary in alleged instances of bullying and harassment is based on three considerations: intensity, persistence, and specificity.

Intensity denotes the graveness of the comment in reference to the target's life or health. Persistence captures the frequency of comments directed towards the target, and specificity measures the size of the aggressor's intended audience and whether the targeted person or group is easily identifiable. This framework can be visualized on a three-dimensional axis. In this chart, intensity is shown on the x-axis, specificity on the y axis, and persistence on the z-axis. These factors are all balanced against one another to determine whether content should remain on Patio's platform. Figure 4 on the following page shows the estimated range of permitted content, denoted by the grey pyramid. Content falling within the grey pyramid is permissible whereas content extending beyond the grey pyramid is a violation of the policy.

Content moderation decisions require considering all three criteria. For example, repeated comments of low intensity and medium specificity can still justify removal, just as a single comment of medium intensity and high specificity encourages moderation action.

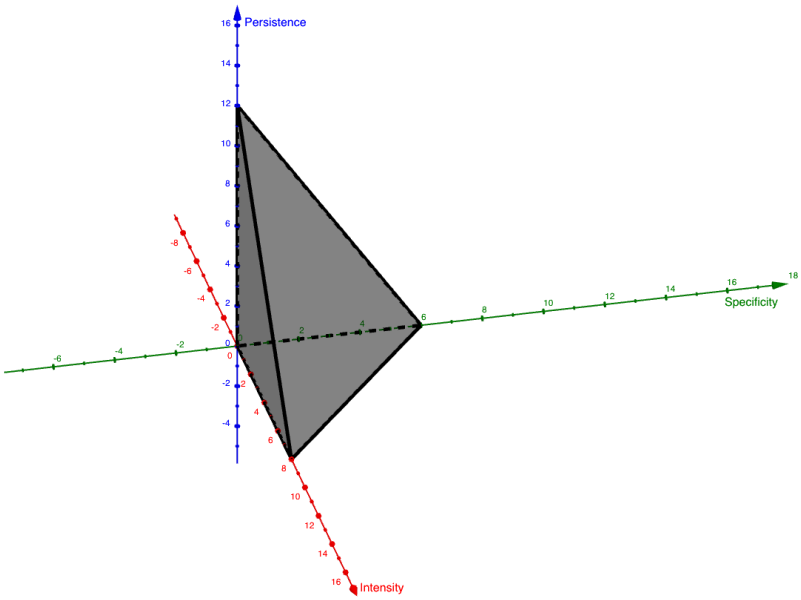

Figure 4: Permitted Content in Tripartite Framework

That said, a comment with low specificity and low intensity would need to be reposted more often than a comment of high specificity with low persistence to justify platform action. This is because direct threats against a singular person are more objectionable than a pattern of aggressive comments directed at a general group. This explains why the threshold for removal is higher on the persistence axis, lower on the intensity axis, and lowest on the specificity axis.

Subsection B.2: Applications to Misinformation

The framework articulated here can guide content moderators in most instances of bullying and harassment. Anti-vaccination content is complicated by its close ties to disinformation campaigns that have potential to prolong a pandemic that has claimed over four million lives worldwide.

Patio has a strong prerogative against misinformation. That said, users who express vaccine hesitancy or share misinformation do not forfeit their protection from bullying and harassment. This framework can be applied to distinguish legitimate social sanction and expressions of disapproval, which are protected by the community guidelines, and prohibited bullying and abuse. Table 5 on the next page explores comments directed at anti-vaxxers in the context of Patio's framework, from higher to lower importance.

It is important to note that each of these reports should be taken seriously in their own right. The purpose of this table is to show the relative priority of each report according to the characteristics of intensity, persistence, and specificity.

Table 5: Examples of Anti-Vaxxer Comments

	Intensity	**Persistence**	**Specificity**
Highest importance	"I'm going to break every anti-vaxxer's arm at school"	Mentioned frequently, several times a day	"[specific user] won't get vaxxed and is living in [dorm]. Stay away from him."
Higher importance	"I want every anti-vaxxer to die"	Mentioned frequently, once a day	"The Republican Club won't get vaccinated. Screw them"
Medium importance	"People who don't get vaccinated deserve to be killed"	Mentioned semi-frequently, a few times a week	"I met this guy in the dining hall who wasn't vaccinated."
Lower importance	"I hope people who don't get vaccinated get COVID"	Mentioned more than once	"I hear there are a few students who didn't get vaccinated"
Lowest importance	"I can't believe these anti-vaxxers, they deserve whatever is coming for them"	Mentioned once	"Anti-vaxxers are a huge threat to society."

Appendix C: Survey Protocol

Subsection C.1: Introduction

Thank you for agreeing to participate in Patio's experiment regarding content moderation. Please be assured that your participation is entirely voluntary and that you may withdraw at any time. In sum, we expect your participation to take 20 minutes or less. We ask that you complete this survey on a computer, as it is not optimized for mobile phones.

The purpose of this survey is to test the efficacy of our content moderation policy prohibiting bullying as it relates to vaccine hesitancy and disinformation. This survey contains six sections. In Section 1, you will be asked for basic information about your age, gender, work experience, and COVID-19 vaccination status. In Section 2, you will be asked to read our publicly available community guidelines or internal policy explanation. In Section 3, you will read a short excerpt of an authentic user conversation. In Section 4, you will be asked to evaluate which messages, if any, violate the community guidelines. In Section 5, you will be asked to read a second user conversation. In Section 6, you will again be asked to assess whether any messages violate our community guidelines.

Please note that by agreeing to participate in this study, you consent for us to use your anonymous responses as part of research that may reach publication. If you have any questions or encounter any difficulties completing this survey, please email [redacted].

Subsection C.2: Participant Information

This is to collect basic information about you to ensure the statistical integrity of our study. Please know that your answers are anonymous.

Question	Response Option
How old are you?	Short answer text
What is your gender?	Multiple choice • Male • Female • Prefer not to answer • Other
Have you been vaccinated against COVID-19?	Multiple choice • Yes • No
Do you work at Patio?	Multiple choice • Yes • No
Have you ever worked in the field of content moderation?	Multiple choice • Yes • No

Subsection C.3: Read Community Guidelines or Policy Explanation

Please read Patio's community guidelines or internal policy explanation. *Note that participants were only provided the resources assigned to them: either the policy explanation or community guidelines, not both.*

Question	Response Option
Were you able to read Patio's community guidelines/internal policy explanation?	Multiple choice • Yes • No

Subsection C.4: Read Case 1

This is an anonymized version of a real conversation on Patio regarding vaccination against the COVID-19 virus.

Question	Response Option
Were you able to access the case?	Multiple choice • Yes • No

Subsection C.5: Responses-Case 1

Please answer the questions below.

Question	Response Option
Do you think any of these messages violated Patio's community guidelines?	Multiple choice • Yes • No
If so, which ones?	Short answer text
How confident are you in your assessment of messages that either did or did not violate Patio's community guidelines?	Ordinal scale • 1 (Not at all confident) • 2 • 3 • 4 • 5 (Very confident)
Additional notes or explanation	Long answer text

Subsection C.6: Read Case 2

This is an anonymized version of a real conversation on Patio regarding vaccination against the COVID-19 virus.

Question	Response Option
Were you able to access the case?	Multiple choice • Yes • No

Subsection C.7: Responses-Case 2

Please answer the questions again.

Question	Response Option
Do you think any of these messages violated Patio's community guidelines?	Multiple choice • Yes • No
If so, which ones?	Short answer text
How confident are you in your assessment of messages that either did or did not violate Patio's community guidelines?	Ordinal scale • 1 (Not at all confident) • 2 • 3 • 4 • 5 (Very confident)
Additional notes or explanation	Long answer text