frontiers in Human Dynamics

BRIEF RESEARCH REPORT
published: 19 February 2021
doi: 10.3389/fhumd.2021.626409

Commercial Versus Volunteer: Comparing User Perceptions of Toxicity and Transparency in Content Moderation Across Social Media Platforms

Christine L. Cook, Aashka Patel and Donghee Yvette Wohn*

Informatics Department, Social Interaction Lab, New Jersey Institute of Technology, Newark, NJ, United States

Content moderation is a critical service performed by a variety of people on social media, protecting users from offensive or harmful content by reviewing and removing either the content or the perpetrator. These moderators fall into one of two categories: employees or volunteers. Prior research has suggested that there are differences in the effectiveness of these two types of moderators, with the more transparent user-based moderation being useful for educating users. However, direct comparisons between commercially-moderated and user-moderated platforms are rare, and apart from the difference in transparency, we still know little about what other disparities in user experience these two moderator types may create. To explore this, we conducted cross-platform surveys of over 900 users of commercially-moderated (Facebook, Instagram, Twitter, and YouTube) and user-moderated (Reddit and Twitch) social media platforms. Our results indicated that although user-moderated platforms did seem to be more transparent than commercially-moderated ones, this did not lead to user-moderated platforms being perceived as less toxic. In addition, commercially-moderated platform users want companies to take more responsibility for content moderation than they currently do, while user-moderated platform users want designated moderators and those who post on the site to take more responsibility. Across platforms, users seem to feel powerless and want to be taken care of when it comes to content moderation as opposed to engaging themselves.

Keywords: Content moderation, Survey Methodology, quantitative research, Social network sites (SNSs), Human Computer Interaction (HCI)

OPEN ACCESS

Edited by:
Wojciech Marek Kulesza,
University of Social Sciences and Humanities, Poland

Reviewed by:
Nathan Schneider,
University of Colorado Boulder, United States
Marc Esteve Del Valle,
University of Groningen, Netherlands

****Correspondence:***
Christine L. Cook
christinelcook@outlook.com

Specialty section:
This article was submitted to Social Networks, a section of the journal Frontiers in Human Dynamics

Received: *05 November 2020*
Accepted: *13 January 2021*
Published: *19 February 2021*

Citation:
Cook CL, Patel A and Wohn DY (2021) Commercial Versus Volunteer: Comparing User Perceptions of Toxicity and Transparency in Content Moderation Across Social Media Platforms. Front. Hum. Dyn 3:626409. doi: 10.3389/fhumd.2021.626409

INTRODUCTION

Although trolling and toxicity online are both hot topics of conversation in the media (Brogunier, 2019; Escalante, 2019; Samson, 2019) and academia alike (Herring et al., 2002; Shachaf and Hara, 2010; Cook et al., 2018) far less-discussed is our current system of dealing with this problem: content moderation. Content moderators are the average netizen's first line of defense, reviewing posts to ensure that they do not violate the website's terms and policies (Roberts, 2016; Carmi, 2019; Wohn, 2019). Though all content moderators work to protect their website's users (Grimmelmann, 2015), the way these protectors are treated and managed differs considerably across platforms (Gillespie, 2019; Seering et al., 2019; Squirrell, 2019; Tyler et al., 2019). Arguably the most significant difference across platforms is the type of content moderators: commercial (contracted) moderators or user volunteers. Though these moderators can use many of the same sorts of tactics—deleting offending posts one by one (Squirrell, 2019), or using more broad measures such as Twitter's blocklists (Jhaver et al., 2018)—the two types of moderation can also vary considerably in terms of how moderators are appointed, who comes up with the rules/guidelines for appropriate/inappropriate content (platform-wide company policy vs. user-policed communities), how transparent they are when enforcing the rules, and how moderators deal with the toxic content and/or offenders (Jhaver et al., 2019; Suzor et al., 2019). Moreover, we know little of how users perceive content moderation and its effectiveness across these varied platforms.

Although there is plenty of work on content moderation, there is scant work in the way of content moderation *theory*. However, the work of Grimmelmann (2015) and Roberts (2016) gives us a framework to work with when describing content moderation practices. Apart from the fact that some moderators are contracted employees and others are user volunteers, there are two other key concepts that can be used to differentiate commercial and volunteer moderation: their centrality and their transparency. Centralized social media platforms, like Facebook, have a head that makes all content moderation rules and decisions—namely, the corporation itself. Contracted content moderators make their decisions based on the corporation's rules and regulations, while on decentralized platforms like Reddit, the rules are both created and enforced by the users, and can differ between the individual communities on the platform. Transparency is a closely-related concept, as it refers to how much of the content moderation decisions are communicated to users. Extant literature would suggest that user-moderated platforms are much more transparent than commercially-moderated platforms (Suzor et al., 2019), but there has yet to be a study that examines how this core difference impacts user perceptions: which type of moderation is preferred, and why. Content moderation is a field of research that is quickly accelerating, with new research being published regularly (e.g., Cai and Wohn, 2019; Jhaver et al., 2019; Seering et al., 2019); without knowing what users are thinking of these developments though, researchers risk alienating the very people they aim to serve with their research, creating an ever-growing gap between what is possible through research and the reality users are experiencing in their day-to-day platform engagement.

To explore user perceptions of current social media content moderation practices, thereby closing the gap between researchers' ideas and users' realities, we conducted a large-scale survey of social media users across six major platforms: Facebook (commercially-moderated), Instagram (commercially-moderated), Reddit (user-moderated), Twitter (commercially-moderated), Twitch.tv (user-moderated), and YouTube (commercially-moderated). Through this survey, participants indicated their own social media habits and understanding of current content moderation practices, as well as what they believe best practices should be when it comes to content moderation. We wanted to see what possibilities users were aware of in terms of policy and toxicity prevention, and what they know about what is currently happening on social media, to better understand where users and researchers stand in relation to one another. However, as described by Grimmelmann (2015) and others (e.g., Gillespie, 2019; Samples, 2019), there are multiple groups that are involved in content moderation; therefore, we wanted to check participants' understanding of content moderation at the 1) corporate level, 2) governmental level, and 3) user level. In this way, we could better grasp the current state of user perception, allowing us to better grasp what is missing in terms of user education about content moderation as it moves forward and evolved. More specifically, we were able to explore the following research questions:

TABLE 1 | Users who use multiple platforms divided by primary social media platform of use.

	Facebook	Instagram	Reddit	Twitch	Twitter	YouTube
N	150	152	151	151	149	149
Facebook	150	120	123	110	116	119
F. Groups	60	54	38	41	43	47
Instagram	90	152	98	89	98	99
Pinterest	40	45	47	27	42	50
Reddit	86	82	151	86	90	104
Snapchat	39	38	43	44	37	42
TikTok	33	24	28	41	28	27
Tumblr	16	16	17	23	22	18
Twitch	47	32	46	151	57	54
Twitter	97	102	105	104	149	98
Wikipedia	54	54	63	66	59	70
YouTube	112	117	123	106	116	149
Other	1	1	0	2	0	1

Note. *F. Groups, Facebook Groups. All "Other" categories were either LinkedIn or Slorum.*

RQ1) How does the type of content moderator—paid worker or user volunteer—influence user perceptions of moderation effectiveness?
RQ2) How familiar are users with current content moderation practices in social media?
RQ3) What do users believe should be done when it comes to content moderation on their favorite social media platforms?

METHODS

Participants and Design

To gather user perceptions regarding content moderation practices and social media, we conducted online surveys with participants recruited through Amazon's Mechanical Turk platform. Recruitment of participants was limited to the United States to ensure that all participants were referring to the same government when questioned about governmental influence in social media. Six parallel surveys were created for six different social media platforms: Facebook/Facebook Groups[1], Instagram, Reddit, Twitch, Twitter, and YouTube.

[1] Perceived positivity or negativity and confidence in knowledge of content moderation practices were not assessed separately for Facebook Groups, as they were considered part of Facebook in the survey.

For each survey, 150 participants were recruited on Mechanical Turk (exact numbers of responses per survey, as well as social media usage statistics per participant, are presented in **Table 1**) per survey. Of the original 1,071 participants, 62 did not complete the survey in its entirety, 103 were disqualified for not being users of the platform in question, and four were disqualified for failing an attention check question (participants were asked to select YDB from a series of three-letter combinations).

Our final 902 participants were aged 18 to 71 (M = 35.31, SD = 10.51) and mostly identified as men (524, 58.1%). Another 313 participants identified as female (34.7%), and eight participants identified as non-binary (0.9%); 57 chose to not disclose their gender identification. The majority of these participants were Caucasian (637, 70.6%), followed by Latino/Hispanic (69, 7.6%), then African American (54, 6.0%), Asian American (38, 4.2%), Native American (8, 0.9%), Middle Eastern (3, 0.3%), or mixed (15, 1.7%). Seventy-eight participants chose to not disclose their race.

Each survey asked a series of multiple-choice questions about toxic content on that social media platform (e.g., "How positive or negative do you perceive [social media platform] to be?", measured on a 5-point Likert scale). Because the surveys were utilized to perform a cross-platform analysis of user perceptions of effective strategies for managing toxicity, each survey asked the same questions for which the answer options were minimally modified to fit the social media platform's affordances. For example, of the platforms surveyed, only Reddit has administrators who are involved in content moderation; thus, although Redditors were asked about how involved administrators (admins) should be in content moderation, none of the other platforms were. On average, participants took 9 min to complete a survey. If the participant had then successfully completed the survey, they were compensated $1.50 (USD) for their time.

Procedure

First, participants were asked the frequency of utilizing the features of the particular social media platform in question and the length of time they have used the platform. They were also asked how confident they were in their knowledge of content moderation practices on the platform, and how positive or negative they perceive the platform to be, negative referring here to toxicity. Following these questions, they were asked who they believe is currently responsible for deciding the appropriateness of a post or comment and who they believe should be responsible for deciding the appropriateness of a post or comment. Specific answer options varied based on the affordances of the platform, but they all fell into one of five categories: the person who posted the post or comment (poster), the intended audience of the post or comment (audience), all users of the platform (users), volunteer moderators (moderators), or commercial content moderators (company). They were also asked how involved they believe the government should be in managing platform-based content, and what initiatives the government should or should not be taking in regards to companies and content moderation.

After expressing their opinions regarding content moderation practices on their platform of choice, they were asked about how users can combat toxicity themselves: "How effective do you think are the following strategies in terms of getting rid of toxicity?" Each survey had different individual strategies based on the particular platform that were grouped (minimum of three strategies per group) into the following categories: blocking, shaming, education/communication, humor, ignoring, reporting, and encouraging positivity. A previous study identified these categories as being the most prominent and effective in handling online harassment (Cai and Wohn, 2019). A full list of questions analyzed in the present study is available in **Supplementary Material**.

RESULTS

All analyses were performed using RStudio (RStudio Team, 2020). Packages used were *dplyr* (Wickham et al., 2020), *questionr* (Barnier et al., 2018), and *sjstats* (Lüdecke, 2020).

User Perception of Platforms' Efficacy in Dealing With Toxicity

In terms of our participants' perception of how toxic they perceived their social media platform of choice to be, there were no significant differences between commercially-moderated (Facebook, Instagram, Twitter, YouTube; M = 0.67, SD = 0.96), and user-moderated (Facebook Groups, Twitch, Reddit; M = *0.79*, SD = *0.89*) platforms, F (1,900) = 3.17, p = 0.07. However, we also examined how confident participants were in their knowledge of content moderation practices on their social media platform of choice and found significant differences between the commercially-moderated and user-moderated platforms (F (1,900) = 4.70, p = 0.03, η^2 = 0.01), with participants reporting generally feeling more confident in their knowledge of content moderation run by volunteers (M = 3.30, SD = 1.23) than by professionals (M = 3.11, SD = 1.29). This is logical, as, in sites run by volunteer moderators, the moderation is visible to users, while it is hidden in commercial moderation situations.

Allocation of Responsibility in Dealing With Toxicity

Participants' perceptions regarding how commercially-moderated and user-moderated platforms currently allocate moderation responsibility across the parties involved for both posts and comments are presented in **Table 2**.

To determine if there are significant differences between how users believe responsibility is *currently* being allocated vs. how they believe it *should* be allocated between the aforementioned parties, we ran a series of t-tests. On commercially-moderated platforms, there are no designated moderators, while on user-moderated platforms, the entirety of the user-based is encompassed by posters, audience, and moderators. The t-test results are as follows. For posts on commercially-moderated

TABLE 2 | Mean scores for responsibility allocation for content moderation according to moderation type.

Mod. Type	Cont. Type	Posters		Audience		Moderators		Users		Company	
		Cur	Ideal	Cur	Ideal	Cur	Ideal	Cur	Ideal	Cur	Ideal
Commercial	Posts	3.16	3.20	2.18	2.13	NA	NA	2.13	2.06	3.06	3.24
	Comm	3.02	3.20	2.41	2.08	NA	NA	2.36	2.09	3.02	3.19
Users	Posts	3.08	2.93	2.46	2.16	2.70	2.70	NA	NA	2.88	3.01
	Comm	2.91	2.82	2.60	2.37	2.84	2.92	NA	NA	2.92	2.99

Note. *On commercially-moderated platforms, there are no designated moderators, while on user-moderated platforms, the entirety of the user-based is encompassed by posters, audience, and moderators. Mod. Type, Type of Moderation; Cont. Type, Type of Content; Cur., Current; Comm., Comments.*

platforms, t-test results were only significant for the company at 3.28 ($p \leq 0.001$). In terms of comments on these websites, however, results were significant for posters is ($t = -2.97$, $p \leq 0.001$), audience (t = 4.48, $p \leq 0.001$), users (t = 3.56, $p \leq 0.001$), and company ($t = -2.60$, $p \leq 0.01$). For posts on user-moderated platforms, t-test results were only significant for posters ($t = 1.97$, $p \leq 0.05$) and audience ($t = 3.50$, $p \leq 0.001$). In terms of comments on these sites, t-test results were only significant for audience ($t = 3.21$, $p \leq 0.001$) and company ($t = -0.88$, $p \leq 0.01$).

In general, it would appear that on commercially-moderated platforms, users want the company to take more responsibility than they currently are for content moderation of posts. However, when it comes to comments, participants believe both the poster and company should be held more accountable for content moderation, and the burden on the audience and users should be reduced. On user-moderated platforms, a different picture emerges. Participants using these platforms seem generally satisfied with how much responsibility both moderators and the company have in terms of content moderation, both when it comes to posts and comments. That said, they want to reduce the responsibility put on both posters and their audience when it comes to moderating posts, and reduce the responsibility put on the audience when it comes to moderating comments.

However, this still leaves us with the question of where users on these platforms believe this extra responsibility taken from posters and their audience should be allocated, if not to the company or the designated moderators. To this end, we also investigated how involved participants believed the government should be in content moderation, and in what kinds of initiatives they should be investing their time. These initiatives are, in order of popularity among our participants, 1) collecting data about social media practices, 2) creating and maintaining a list of best practices concerning content moderation on social media, 3) monitoring companies to ensure that they adhere to their content moderation responsibilities, 4) contacting offenders on social media to enforce policies, and 5) directly monitoring social media content.

The means and standard deviations for these proposed initiatives for participants who primarily use commercially-moderated or user-moderated platforms are as follows. For commercially-moderated platforms, for 1) collecting data the mean was 3.38 and the standard deviation 1.30, 2) best practices the mean was 2.98 and the standard deviation was 1.34, 3) monitoring companies the mean was 2.76 and the standard deviation 1.49, 4) contacting offenders the mean was 2.63 and the standard deviation 1.49, and 5) monitoring content the mean was 2.52 and the standard deviation was 1.53. For user-moderated platforms, for 1) collecting data the mean was 3.30 and the standard deviation 1.32, 2) best practices the mean was 2.94 and the standard deviation was 1.38, 3) monitoring companies the mean was 2.82 and the standard deviation 1.48, 4) contacting offenders the mean was 2.66 and the standard deviation 1.50, and 5) monitoring content the mean was 2.58 and the standard deviation was 1.55. Facebook Groups is considered a part of Facebook for the purpose of these scores.

We first noticed that the order of importance/popularity of these different initiatives remains the same, regardless of how the platform is moderated: participants seem to believe that the government should be more involved in collecting data (e.g., how often companies are actively punishing offenders, how many offensive posts are caught per month) than in the direct monitoring of social media content. It is also worth noting that the only initiative that scored an average over 3.00 was data collection, suggesting that most users do not seem to want the government to be too involved in content moderation.

User-Based Toxicity Interventions and Efficacy

At the outset, we wanted to discover how often our participants engaged in toxicity management strategies such as reporting posts or offenders and messaging moderators or website administrators. However, there are many factors that could influence the frequency of participants' engagement in toxicity management: participants' age, beliefs about how content moderation is currently being handled, beliefs about how it should be handled, and how intense a social media user they are, to name a few. Nonetheless, our primary interest was in how often users on commercially-moderated platforms vs. users on user-moderated platforms engage in toxicity management. An ANOVA revealed that there was no such significant difference (F (1,900) = 2.18, p = 0.14), but we still wanted to investigate if the amount of time spent by users engaging in toxicity management was predicted by the same variables on both commercially-moderated and user-moderated platforms. To explore this, we tested two models, each predicting the amount of time spent on toxicity management: one on commercially-moderated platform users, and one on user-moderated platform users. Our results are presented in **Table 3**.

Generally, it would seem that on both commercially-moderated and user-moderated platforms, how confident a

TABLE 3 | Regressions predicting the amount of time spent by participants on toxicity management.

	Commercially-moderated			**User-moderated**		
	***F* (24,402) = 34.29, R^2 = 0.65**			***F* (24,116) = 19.58, R^2 = 0.76**		
	B	**Σ**	***P***	**B**	**σ**	***P***
Age	−0.01	0.004	0.42	−0.04	0.01	0.29
Gender (male to female)	−0.06	0.08	0.02	−0.01	0.13	0.56
Gender (male to non-binary)	−1.06	0.47	0.49	0.07	0.71	0.89
Intensity of social media usage	0.08	0.15	0.03	−0.38	0.27	0.17
Confidence in knowledge of content moderation practices	0.13	0.04	< 0.001	0.19	0.06	0.003
Positive/Negative perception of the social media platform	−0.06	0.05	0.20	−0.13	0.08	0.11
Passive consumption	0.13	0.06	0.05	0.28	0.12	0.02
Active consumption	0.22	0.06	< 0.001	0.59	0.12	< 0.001
Beliefs regarding current levels of engagement in moderation practices (by content type) by:						
Poster (posts)	−0.02	0.06	0.70	−0.16	0.09	0.07
Audience (posts)	0.15	0.06	0.02	0.09	0.08	0.25
All users (posts)	0.16	0.06	0.01	*NA*	*NA*	*NA*
Moderators (posts)	*NA*	*NA*	*NA*	−0.08	0.12	0.48
Company (posts)	−0.005	0.06	0.94	0.03	0.12	0.81
Poster (comments)	0.01	0.05	0.86	0.09	0.08	0.27
Audience (comments)	0.04	0.05	0.42	−0.02	0.08	0.83
Users (comments)	0.004	0.05	0.94	*NA*	*NA*	*NA*
Moderators (comments)	*NA*	*NA*	*NA*	−0.11	0.11	0.31
Company (comments)	0.07	0.05	0.20	0.09	0.12	0.48
Beliefs regarding ideal levels of engagement in moderation practices (by content type) by:						
Poster (posts)	−0.08	0.06	0.23	0.07	0.08	0.36
Audience (ideal - posts)	−0.02	0.07	0.76	0.15	0.07	0.05
All users (posts)	0.24	0.06	< 0.001	*NA*	*NA*	*NA*
Moderators (ideal - posts)	*NA*	*NA*	*NA*	−0.04	0.14	0.77
Company (posts)	−0.02	0.06	0.72	−0.04	0.10	0.68
Poster (comments)	0.01	0.07	0.91	−0.01	0.08	0.94
Audience (comments)	0.06	0.06	0.34	0.17	0.08	0.02
All users (comments)	0.11	0.06	0.09	*NA*	*NA*	*NA*
Moderators (comments)	*NA*	*NA*	*NA*	−0.24	0.14	0.09
Company (comments)	−0.02	0.05	0.65	0.08	0.09	0.38

Note. *On commercially-moderated platforms, there are no designated moderators, while on user-moderated platforms, the entirety of the user-based is encompassed by posters, audience, and moderators. Significant results are bolded.*

user is in the platform's content moderation practices has a significant influence on how much toxicity management users engage in, with more confident users engaging in more toxicity management. Exactly how much a user engages with the platform also helps determine how much toxicity management a user performs, whether that be doing things like actively posting and commenting (active consumption), or simply browsing and clicking "like" every now and then (passive consumption).

From here, the results begin to diverge more significantly between commercially- and user-moderated platforms. On commercially-moderated platforms, more intense users, i.e., those who use the most functions on social media the most often, also engage in more toxicity management, further corroborating our results with active and passive consumption. Gender also plays a role in the amount of toxicity management users engage in, with men performing more than women, on average. None of these effects were significant on user-moderated platforms. In addition, participants' beliefs about current levels of engagement by both the intended audience of posts and all users generally when it comes to posts both predicted toxicity management levels on commercially-moderated platforms. The more involved participants believed that the intended audience and users more generally are, the more toxicity management they themselves engaged in, while this was not the case for user-moderated platforms. Their ideal amount of content moderation engagement in terms of general users of the platform also impacted toxicity management levels on commercially-moderated platforms, with more engagement ideally leading to more toxicity management by participants. On user-moderated platforms, beliefs about current moderation engagement had no impact on actual amounts of toxicity management by participants; however, beliefs about ideal content moderation engagement did. The more engagement the participant in question believed the audience should engage in, be it for posts of comments, the more toxicity management they themselves performed. In sum, it would seem that if our participants believe others should be held accountable for toxicity, they also hold themselves accountable and try to do their part to clean up social media.

However, we also wanted to see what kind of user-based interventions our participants thought were effective in this clean-up effort. Is that time spent reporting offenders more effective on commercially-moderated or user-moderated websites, for instance? The scores for each intervention type evaluated are presented in **Table 4**.

TABLE 4 | Perceptions of user strategy effectiveness in dealing with toxicity.

	Commercially-moderated		User-moderated	
	M	***SD***	***M***	***SD***
Blocking	2.91	0.68	2.93	0.68
Reporting	2.84	0.75	2.80	0.71
Positivity	2.59	0.95	2.57	0.86
Ignoring	2.48	0.96	2.48	0.93
Edu/Comm	2.21	0.82	2.19	0.91
Humor	1.96	0.95	2.05	0.95
Shaming	1.89	0.95	2.07	0.89

Note. *Edu/Comm, Education and Communication. Positivity refers to the encouragement or enacting of prosocial behavior on the platform.*

Once more, we see that the pattern or order of effectiveness barely differs between commercially-moderated and user-moderated platforms; apart from humor being perceived as slightly more effective than shaming on commercially-moderated platforms, and the reverse being true on user-moderated platforms, there are no differences. Blocking and reporting offenders are consistently perceived as being the most effective user-based toxicity management strategies, while humor and shaming are consistently perceived as being the least effective. However, to further explore the differences between users of commercially-moderated and user-moderated platforms, we also ran several ANOVAs to see if the type of moderation made any difference to the perception of user effectiveness. All but one of these were insignificant [*F*s (1,900) ≤ 1.71, *p*s > 0.19]: shaming was perceived as slightly more effective on user-moderated platforms than on commercial platforms, F (1,900) = 7.76, $p = 0.01$, $\eta^2 = 0.01$. However, it is important to note that none of these mean values surpassed 3.00, suggesting that users generally find user-based toxicity management to be largely ineffective.

DISCUSSION

In terms of our first– how does the type of content moderator, employee or user volunteer, influence user perceptions of moderation effectiveness—and second—how familiar are users with current content moderation practices in social media—research questions, we found mixed results. Although participants did not see a major difference in toxicity levels between commercially-moderated or user-moderated platforms, they did feel significantly more confident in their understanding of content moderation practices on user-moderated platforms. Participants who were more confident in this knowledge also reported engaging in more toxicity management behaviors themselves. This is in line with earlier studies that suggest moderation transparency is critical to ensuring users know how to behave online (West, 2018; Jhaver et al., 2019; Suzor et al., 2019). However, it seems as though participants did not believe their own efforts were effective, as toxicity levels remained the same, irrespective of the type of moderation. This is an interesting result, as existing content moderation work would suggest that at least some user interventions *are* effective toxicity deterrents (Leavitt and Robinson, 2017). Platforms may want to consider employing new strategies when it comes to educating users regarding effective user-based toxicity interventions, as there appears to be a disconnect between the user perception and the social media reality.

This idea of user ineffectiveness is also reflected in the results pertaining to our third research question: what do users believe should be done when it comes to content moderation on their favorite social media platforms? Although blocking and reporting were perceived as being the most effective strategies a user could employ from those listed (Cai and Wohn, 2019), none of the strategies were considered very effective overall. Instead, while on commercially-moderated platforms, participants wanted the company to take more responsibility for moderation and remove the burden from users, participants on user-moderated platforms wanted the moderators to have the most responsibility and the posters to also take responsibility for what they put online. This is interesting in light of the importance of transparency (Roberts, 2016; West, 2018; Jhaver et al., 2019; Suzor et al., 2019); though our participants were often heavy users of social media and engaged in significant amounts of toxicity management, it would seem that they would rather content moderation be a separate job. This would suggest that though transparency is important and effective (Grimmelmann, 2015; Carmi, 2019; Jhaver et al., 2019; Tyler et al., 2019), it is not desirable; social media users seem to appreciate having a big brother figure—be that a corporation, a more responsible user, or the government—take care of the details without individual users having to worry about making these decisions. This could, however, be due to social media users being unaware of the full extent of the possibilities regarding their own involvement. Though not necessarily common practice today, social media users could someday be involved in the creation of the policies that govern them, or perhaps "content moderator" could become an elected position. Future studies into content moderation and its effectiveness should still be careful to take the user's perspective into consideration, but not let that limit how social media grows and evolves over time.

Of course, this study can only be interpreted in the social media context; it is entirely possible that these patterns do not hold in other contexts, such as online games or more specialized online fora. Future studies could investigate these other platforms to see if the desire for governmental or corporate involvement in content moderation differs elsewhere. In addition, this survey was conducted among Americans, albeit varied in racial identity; future studies should seek to test these results in other cultures to see what is generalizable and what is not. However, that said, research concerning content moderation among Americans typically brings up questions of free speech (Wohn et al., 2017; Samples, 2019), while our results would suggest that our American participants actually appreciate the caretaker role of government and corporation. There is still considerable room for scholarship concerning how we balance power, responsibility, and freedom on the Internet across all cultures and platforms.

DATA AVAILABILITY STATEMENT

The raw data supporting the conclusions of this article will be made available by the authors, without undue reservation.

ETHICS STATEMENT

The studies involving human participants were reviewed and approved by the Institutional Review Board of the New Jersey Institute of Technology. The patients/participants provided their written informed consent to participate in this study.

AUTHOR CONTRIBUTIONS

CC wrote the introduction, the procedure section of the methods, results, and the majority of the discussion. CC conducted all quantitative analysis of the data and contributed to revisions. AP wrote the participants/design section of the methods and parts of the discussion section as well as contributed to revisions. AP was responsible for creating all submission documents. DW was the project supervisor who conceptualized the study and provided critical feedback during the process and contributed to revisions. All authors participated in designing the six different surveys.

FUNDING

This study was funded by the National Science Foundation (NSF) award 1841354.

SUPPLEMENTARY MATERIAL

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fhumd.2021.626409/full#supplementary-material.

REFERENCES

Barnier, J., Briatte, F., and Larmarange, J. (2018). questionr: functions to make surveys processing easier. R package version 0.7.0. Available at: https://CRAN.R-project.org/package=questionr.

Brogunier, T. (2019). 4 reasons why social media has become so toxic and what to look for next. Available at: https://www.entrepreneur.com/article/328749 (Accessed August 17, 2020).

Cai, J., and Wohn, D. Y. (2019). "What are effective strategies of handling harassment on twitch?," in CSCW'19: conference companion publication of the 2019 on computer supported cooperative work and social computing, Austin, TX, November 2019, 166–170. Available at: https://doi.org/10.1145/3311957.3359478.

Carmi, E. (2019). The hidden listeners: regulating the line from telephone operators to content moderators. *Int. J. Commun.* 13, 440–458. Available at: https://ijoc.org/index.php/ijoc/article/view/8588/2540 (Accessed August 17, 2020).

Cook, C., Schaafsma, J., and Antheunis, M. (2018). Under the bridge: an in-depth examination of online trolling in the gaming context. *New Media Soc.* 20, 3323–3340. doi:10.1177/1461444817748578

Escalante, A. (2019). Is social media toxic to your teen's mental health?. Available at: https://www.psychologytoday.com/us/blog/shouldstorm/201909/is-social-media-toxic-your-teens-mental-health (Accessed August 17, 2020).

Gillespie, T. (2019). *Custodians of the internet: platforms, content moderation, and the hidden decisions that shape social media*. London, United Kingdom: Yale University Press.

Grimmelmann, J. (2015). The virtues of moderation. *Yale J. Law Tech.* 17, 42–109. Available at: https://digitalcommons.law.yale.edu/yjolt/vol17/iss1/2/#:~:text=Yale%20Journal%20of%20Law%20and%20Technology&text=It%20breaks%20down%20the%20basic,sterility%20of%20too%20much%20control (Accessed August 17, 2020). doi:10.31228/osf.io/qwxf5

Herring, S., Job-Sluder, K., Scheckler, R., and Barab, S. A. (2002). Searching for safety online: managing "trolling" in a feminist forum. In *Inf. Soc.* 18, 371–384. doi:10.1080/01972240290108186

Jhaver, S., Bruckman, A., and Gilbert, E. (2019). Does transparency in moderation really matter?: user behavior after content removal explanations on reddit. *Proc. ACM Hum. Comput. Interact.* 3, 27. doi:10.1145/3359252

Jhaver, S., Ghoshal, S., Bruckman, A., and Gilbert, E. (2018). Online harassment and content moderation: the case of blocklists. *ACM Trans. Comput. Hum. Interact.* 25. doi:10.1145/3185593

Lüdecke, D. (2020). *sjstats: statistical functions for regression models*. R package version 0.17.9. Available at: https://CRAN.R-project.org/package=sjstats.

Leavitt, A., and Robinson, J. J. (2017). "The role of information visibility in network gatekeeping: information aggregation on Reddit during crisis events," in Proceedings of the 2017 ACM conference on computer supported cooperative work and social computing, 1246–1261.

Roberts, S. T. (2016). Commercial content moderation: digital laborers' dirty work. In *Media studies publications*, 12. Available at: https://ir.lib.uwo.ca/commpub/12 (Accessed August 17, 2020).

RStudio Team (2020). *RStudio: Integrated Development for R*. Boston, MA:RStudio, PBC. Available at: http://www.rstudio.com/16.

Samples, J. (2019). Why the government should not regulate content moderation of social media. *Policy Anal.*, 32, 2019 Number 865. Available at: https://ssrn.com/abstract=3502843 (Accessed August 17, 2020).

Samson, D. (2019). Poll declares "like" button as one of the most toxic social media features. Available at: https://www.techtimes.com/articles/245180/20190830/poll-declares-like-button-as-one-of-the-most-toxic-social-media-features.htm18 (Accessed August 17, 2020).

Seering, J., Wang, T., Yoon, J., and Kaufman, G. (2019). Moderator engagement and community development in the age of algorithms, *New Media Soc.* 21 (2), 1417–1443. doi:10.1177/1461444818821316

Shachaf, P., and Hara, N. (2010). Beyond vandalism: wikipedia trolls. *J. Inf. Sci.* 36, 357–370. doi:10.1177/0165551510365390

Squirrell, T. (2019). Platform dialectics: the relationships between volunteer moderators and end users on reddit, *New Media Soc.* 21, 1910–1927. doi:10.1177/1461444819834317

Suzor, N. P., West, S. M., Quodling, A., and York, J. (2019). What do we mean when we talk about transparency? Toward meaningful transparency in commercial content moderation. *Int. J. Commun.* 13, 1526–1543. Available at: https://ijoc.org/index.php/ijoc/article/view/9736 (Accessed August 17, 2020).

Tyler, T., Katsaros, M., Meares, T., and Venkatesh, S. (2019). Social media governance: can social media companies motivate voluntary rule following behavior among their users?. *J. Exp. Criminol.* doi:10.1007/s11292-019-09392-z

West, S. M. (2018). Censored, suspended, shadowbanned: user interpretations of content moderation on social media platforms. *New Media Soc.* 20, 4366–4383. doi:10.1177/1461444818773059

Wickham, H., François, R., Henry, L., and Müller, K. (2020). *dplyr: a grammar of data manipulation*. R package version 0.8.5. Available at: https://CRAN.R-project.org/package=dplyr24.

Wohn, D. Y., Choudhury, M. D., Fiesler, C., Matias, J. N., and Hemphill, L. (2017). How to handle online risks? Discussing content curation and moderation in social media. In CHI'17 extended abstracts, Denver, CO, May 06-11, 2017, 1271–1276. Available at: http://dx.doi.org/10.1145/3027063.3051141.

Wohn, D. Y. (2019). "Volunteer moderators in Twitch micro communities: how they get involved, the roles they play, and the emotional labor they experience," in Proceedings of the 2019 CHI conference on human factors in computing systems, Paper 160, Glasgow, Scotland United Kingdom, May 4–9, 2019 (CHI 2019), 1–13. Available at: https://doi.org/10.1145/3290605.3300390.

Conflict of Interest: The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.