User Opinions on Effective Strategies Against Social Media Toxicity

Aashka Patel
New Jersey Institute of Technology
asp263@njit.edu

Christine L. Cook
New Jersey Institute of Technology
christine.l.cook@njit.edu

Donghee Yvette Wohn
New Jersey Institute of Technology
wohn@njit.edu

Abstract

Existing literature on content moderation rarely identifies strategies social media users believe can be implemented by platforms and other users to adequately manage toxicity and curate a positive environment online. To bridge these gaps, we conducted a survey with 902 users of six different social media platforms to understand their opinions on effective strategies against social media toxicity and for a positive online environment. Participants suggested a range of strategies, both for the platforms and the users to implement. Even though all these strategies are not unprecedented, it is crucial to recognize that currently not all platforms implement or can implement these strategies to encourage less toxicity and more positivity. Overall, participants expect platforms and social media users to do their individual and collective part in managing online toxicity.

1. Introduction

Toxicity online is a major issue, especially as it proliferates on social media and has negative impacts on both individuals and society. Content moderation is a means of handling some of that negativity, but there is no "one method" of content moderation, as all companies have different policies, technologies, and personnel associated with these activities. In this study, we asked U.S. adult social media users of six different platforms (Facebook, Instagram, Twitter, Twitch, YouTube, Reddit) about their opinions on what they thought would be an effective strategy toward getting rid of toxicity online and fostering a positive online environment. The opinions of "average" social media users are important as most social media companies rarely solicit the preference of users.

2. Background

In 2019, Pew Research found that 72% of Americans use social media platforms to exchange information, explore news and media, connect with others, and find entertainment [1]. This is a significant increase from the 42% of Americans that utilized social media at the end of 2009 [1]. As the use of social media rises through the years, the voices of the horrific and toxic aimed at harming society grow with it [2]. Online toxicity can manifest in various forms including hate speech, bullying, trolling, harassment, physical threats, and online stalking [3]. Toxicity can turn from threats and taunts posted on social media platforms to real-world violence [4]. For example, in 2016, interim police Superintendent John Escalante credited the drastic spike in shootings and violence in Chicago to the threats posted on Twitter and Facebook between young gang members [5]. However, online toxicity is not limited to just gang members. In a 2017 survey conducted by Pew Research, it was found that 41% of Americans have been subjected to harassing behavior online personally and 66% have witnessed harassing behavior targeted towards others [6]. Take for example Michele Dauber, a 53-year-old Stanford Law School Professor, who initiated a campaign to recall a controversial judge and people threatened to rape her and cut her throat [7]. To this day, Dauber suffers from panic attacks as a result of this harassment [7]. A survey conducted by the Anti-Defamation League in 2018 also found that 37% of Americans have experienced online hate and harassment in high severity [8]. Sexual harassment, offensive name-calling, purposeful embarrassment, swatting, doxxing, stalking, and physical threats are examples of this [8, 6]. Online abuse is also frequently targeted in nature, as 33% of Americans experience abuse because of their "sexual orientation, religion, race, ethnicity, and gender identity or disability" [9].

Clearly, online toxicity is an issue that is widespread and can result in serious consequences – both virtually

URI: https://hdl.handle.net/10125/70980
978-0-9981331-4-0

and physically. For these reasons, must be addressed with great care and diligence. With the increase in social media usage and toxicity engagement, toxicity management presents to be a challenging problem [10]. Everyday users of social media platforms, or the people who are frequently exposed to online toxicity, have varying perspectives on how this toxicity should be dealt with and which strategies are effective in its management [6].

2.1. Content Moderation

A commonly-used method of toxicity management by the biggest social media platforms is content moderation [11]. Content moderation is the "organized practice of screening user-generated content (UGC) posted to Internet sites, social media and other online outlets, in order to determine the appropriateness of the content for a given site, locality, or jurisdiction" [12]. In content moderation, there are three important distinctions to be made in terms of practice: manual versus automated, transparent versus opaque, and centralized versus decentralized [13].

Content moderation can be either manual, meaning it is conducted by human moderators who decide what content should be kept or removed, or automatic, meaning algorithms are responsible for filtering content based on specified patterns [13]. Human moderators are able to be "experts in matters of the site's presumed audience and have cultural knowledge...[and] have expert knowledge of user guidelines and other incredibly detailed platform-level specifics concerning what is and is not allowed" [14]. This can be used alone or in tandem with automated content moderation tools, such as image recognition software, metadata filtering tools, and natural language processing techniques, to identify and remove toxic content [15]. Neither method of moderation is without its flaws. Human moderators like Lester, a past moderator for both YouTube and Twitter who spent nine hours a day deciding if a child's genitals were being touched accidentally or purposefully in images and if a knife slashing a human's throat was a real-life killing or not, are subjected to such horrific content regularly [16]. On the other hand, automated content moderation is not necessarily a replacement for human moderation, as automated tools lack the critical reflection of humans and can falsely remove content that should remain and miss content that should be removed [17, 18, 19]. Clearly, neither method is perfect, but both are useful.

Moreover, content moderation can be done transparently - explaining all decisions to remove content and the reasoning behind those decisions to the public - or opaquely: hiding these decisions from the public eye [13]. With opaque moderation, users have little data available to them about moderation decisions and cannot thoroughly examine the platform's practices [20]. A lack of transparent content moderation decisions can make it so that users lose trust in social media platforms and are less able to comprehend content regulation [20]. According to York and McSherry [18], users have the right to know why their content has been removed or their account suspended; platforms need to release data that includes how many posts and accounts were removed and banned with the reasons why [21]. With transparent moderation being tied to trust and credibility, increased transparency will allow users to better trust their social media platform, and platforms to demonstrate corporate social responsibility [22].

Lastly, moderation can be centralized, with one powerful moderator controlling the decisions, or decentralized, with various moderators making the decisions [13]. Platforms such as Facebook and Twitter use centralized moderation, meaning their harassment and moderation policies are enforced by the platform [23, 24]. Some platform-based moderation practices include hiring moderators to review content, implementing automated tools for detecting and removing toxic content, and banning violators of the platform's policies [25, 17, 26]. On the other hand, platforms such as Reddit and Twitch heavily rely on user-based moderation practices such as utilizing volunteer moderators that govern their communities [27]. Despite the distinctions in moderation practices, the goals of content moderation are the same: "to protect one user from another, or one group from its antagonists, and to remove the offensive, vile, or illegal – as well as to present [the platform's] best face to new users, to their advertisers, and partners, and to the public at large" [23].

A survey conducted by Pew Research found that 79% of American social media users believe that online platforms should be responsible for preventing harassing and toxic behavior on their platforms [6]. However, only 31% of U.S. adults have confidence in social media platforms to be able to determine what content should be removed [28]. The same survey found that the vast majority of Americans, or 62%, consider online harassment to be a major problem. When they were asked how platforms should prevent such behaviors, it was found that 35% of Americans believe that better policies and tools from online platforms are the most effective method in handling these behaviors [6]. Other methods that many Americans believe would be effective in managing online harassment are increasing law enforcement, creating stronger laws, making filtering content on platforms more simple, and making

it easier to report such content [8, 6]. Another study found that educating the toxic person, sympathizing with the toxic person, shaming the toxic person, blocking the toxic person, ignoring the toxic person, humoring the content, and encouraging positivity on the platform are prominent and effective strategies of handling harassment on Twitch [29]. Despite there being differences in what users believe to be effective in managing harassment online, the impact of such toxicity is clearly seen in the change of users' online behavior. Of users who are targets and victims of harassment, 38% have either stopped, reduced, or changed their online activities by deleting apps, avoiding websites, posting less, changing their privacy settings, and applying filters; 18% contacted the platform for help; 15% have taken physical precautions such as taking self-defense classes, avoiding being alone, moving locations, and taking a different commute; and 6% have contacted law enforcement [8]. While these studies give us an idea of how users perceive toxicity and how they personally deal with toxicity, we are not given a clear indication of the specifics of the strategies they believe the platform or *other* users should implement for managing toxicity. Moreover, we are still unsure if and how users believe a non-toxic, or even positive, environment can be curated on social media platforms.

In this study, we report on strategies of toxicity management on social media platforms that users believe to be effective, as well as user perspectives on how a positive environment can be created on social media platforms. This study intends to close the previously mentioned gaps in the existing literature by expanding the scope of our research to various social media platforms and addressing the specifics of what strategies users believe to be effective in toxicity management at the platform-level and the user-level. With this in mind, we developed the following research questions:

RQ1) Are there any strategies that you think are effective in getting rid of toxicity?

RQ2) What helps create a positive environment based on your [social media platform name i.e. YouTube, Facebook, etc.] experience?

Our results will help social media platforms and users understand their current options to manage toxicity and suggest new methods to handle toxicity and create a positive environment online.

3. Methods

To understand the strategies and behaviors users perceive to be effective in handling toxicity and creating a positive environment on social media platforms, we conducted six cross-parallel surveys for users of the following six different social media platforms: YouTube, Facebook/Facebook Groups, Twitter, Instagram, Twitch, and Reddit. These surveys were conducted on Amazon's Mechanical Turk platform with participants recruited from the United States. For each survey, there were approximately 150 participants. Across all surveys, there were 1,071 participants. However, 103 participants were disqualified for not being users of the social media platform the survey asked about. We determined if participants were not users of the social media platform the survey asked about by asking all participants to select the social media platforms they use from a list. If a participant did not select the platform the survey was asking about because they did not use the platform, then they were disqualified from the survey. Additionally, 62 participants left the survey incomplete and 4 participants were removed from analysis for failing our attention check question which asked them to pick "YDB" from a selection of three-letter combinations. Participants were compensated $1.50 (USD).

To gauge how familiar the participants were with the social media platform they were being asked about, we asked them the length of time they have used the platform and their frequency of using certain features of the platform.

For Instagram, 3.2% of participants have been using the platform for 10 or more years, 31.6% for 7 to 9 years, 26.1% for 4 to 6 years, 21.9% for 1 to 3 years, and 5.2% for less than 1 year.

For Twitter, 4.0% of participants have been using the platform for 10 or more years, 16.8% for 7 to 9 years, 30.9% for 4 to 6 years, 28.9% for 1 to 3 years, and 19.5% for less than 1 year.

For Twitch, 10.0% of participants have been using the platform for 10 or more years, 39.1% for 7 to 9 years, 36.4% for 4 to 6 years, 10.6% for 1 to 3 years, and 4.0% for less than 1 year.

For Reddit, 1.3% of participants have been using the platform for 10 or more years, 33.1% for 7 to 9 years, 41.7% for 4 to 6 years, 17.9% for 1 to 3 years, and 6.0% for less than 1 year.

For YouTube, 6.7% of participants have been using the platform for 10 or more years, 14.8% for 7 to 9 years, 28.2% for 4 to 6 years, 49.7% for 1 to 3 years, and 0.7% for less than 1 year.

For Facebook, 50.7% of participants have been using the platform for 10 or more years, 29.3% for 7 to 9 years, 15.3% for 4 to 6 years, 4.0% for 1 to 3 years, and 0.7% for less than 1 year.

The six surveys asked participants for optional demographic information which included gender, age,

and race. Of our 902 participants that completed the surveys in their entirety, 524 identified as men (58.1%), 313 as women (34.7%), 8 as non-binary (0.9%), and 57 did not provide their gender (6.3%). Participants were aged 18 to 71 with the average being 35.31 years and a standard deviation of 10.51 years. The majority of participants were Caucasian (70.6%), followed by Latino/Hispanic (7.6%), African American (6.0%), Asian American (4.2%), Native American (0.9%), Middle Eastern (0.3%), and mixed (1.7%), and 8.7% did not provide their race.

These surveys were conducted for a larger study focused on users' perceptions of content moderation. However, this present study is focused on two open-ended questions we asked the participants in each of the six surveys: RQ1) Are there any other strategies that you think are effective in getting rid of toxicity? and RQ2) What helps create a positive environment based on your [social media platform name] experience?

In our analysis of the data, we followed the grounded theory approach for qualitative research as described by Corbin and Strauss[30]. We grouped data based on no a-priori theory. We utilized the key components of grounded theory: collecting data, identifying concepts, coding our data into themes, and categorizing responses.

Two members of the team went through the participant responses for each of the surveys and highlighted strategies, sorting ideas that were similar while focusing more on strategies that were found to be unique and different from what was already present in literature, as well as responses that were unique in how to create a positive environment. Any responses that were gibberish or deviant (e.g., repeating the same word over and over again) were removed. The responses from RQ1 and RQ2 were then summarized into keywords and concepts that preserved the essence of the original responses.

After this step, we compiled a spreadsheet with the "Survey Source" (which platform the survey pertains to), "Strategies", and "Positive Environment" as the column headings and populated it with our summarized keywords and concepts. One of us went through the summarized responses for RQ1 to create themes and definitions; the responses were then categorized into those themes. Another team member went through the summarized responses for RQ2 to create themes and definitions as well; those responses were then categorized into those themes. To ensure participant responses were categorized into the best-fitting theme, the whole team went through each response and corresponding theme. If there were any disagreement in how a response was categorized, a discussion was had and the proper adjustment to the response's theme was made.

4. Which Strategies are Effective in Dealing with Toxicity?

From the individual responses, we distilled 37 themes across our two research questions: 1) which strategies are effective in dealing with toxicity, and 2) what contributes to a positive social media experience.

In this category, we found four major themes, two of which included several subthemes: user-based interventions (8 subthemes), platform-based interventions (12 subthemes), outside intervention, and no intervention. Even though participants' responses were coded into themes, this categorization was not mutually exclusive. If participants' responses fit into multiple subthemes, they were coded accordingly.

4.1. User-based Interventions

Our first major theme was user-based interventions which include strategies our participants suggested that the user has to employ to deal with online toxicity. There are 54 responses that were categorized into this theme.

4.1.1. Avoidance Our subtheme avoidance is when users avoid certain people or topics on social media that are contentious or controversial. This can include "[not leaving your] profile visible to the public" (P49, Facebook), "just completely turning off the comments" (P54, YouTube), or not using the platform at all (P97, Twitter). The majority of participants' responses in this subtheme found politics and religion to be controversial topics. There were 7 responses were categorized into this subtheme.

4.1.2. Self-Control Self-control is a subtheme defined as users having restraint over their actions on social media platforms. There were 3 responses were categorized into this subtheme. Participant responses suggest that users need to think twice before posting content. P100 (Instagram) described this as "being careful about what kind of content that you post/curate", while P103 (Twitter) emphasized being aware of what you share, and ensuring that it is truthful.

4.1.3. Curating Similar to the concept of avoidance discussed earlier is - curating - which is when users carefully select friends and topics to build an ideal community around themselves. Often this revolves around only friending people you know in person

(P62, Facebook), although P85 (Twitter) described it as follows: "Not interacting with the userbase at large - follow specific people/accounts and kind of create your own insulated bubble". Additionally, users can choose to "turn off [toxic] posts" so that the user does have to see them (P1, Facebook). There were 11 responses were categorized into this subtheme.

4.1.4. Ignoring In alignment with our subthemes of avoidance and curating is ignoring. This subtheme is defined as simply not giving attention to toxic people online and their toxic content. There were 9 instances of this subtheme. One YouTube user suggested an effective strategy is to "not respond to them...Because more are just trying to get a reaction and bully people. If you ignore them, most of the time they just move one because they got bored" (P110). Other participants expressed similar ideas of ignoring toxic people and not feeding into the negativity. For example, an Instagram user wrote, "Don't give the toxic person the attention they are wanting. Ignore it, or report it." (P31).

4.1.5. Self-Care Arguably the most surprising of our subthemes is self-care with 2 responses being categorized into this subtheme. Participants suggested that if people take care of themselves (i.e., hydration, meditation, reasonable amounts of sleep), they will be less likely to be toxic themselves, thereby reducing toxicity levels overall, like P3 (Facebook) who suggested that users need to "pay attention to what [they] feel and need."

4.1.6. Shaming This subtheme of shaming involves addressing the toxic person directly to make them feel bad about the content they posted and even themselves. There were 6 instances of this subtheme. Actions that people take to shame toxic people include: saying mean and rude things back to the toxic person, calling on others to humiliate the toxic person, and purposefully resurfacing the toxic person's old content for others to judge.

4.1.7. Reciprocal Toxicity Another subtheme under the user-based intervention category and similar to our previous subtheme of shaming is reciprocal toxicity: when users respond to toxicity with being toxic themselves. The example our sample gave was doxing, which is revealing private information - presumably a troll's or toxic person's - on the internet without permission. There was only 1 instance of this subtheme.

4.1.8. Reporting Our subtheme of reporting was described frequently by our participants and includes two distinct types: reporting to the platform and reporting to the police. There were 21 instances of this subtheme. Reporting to the platform is defined as reporting content found to be offensive or toxic to the platform for review and removal. A YouTube user finds "Reporting (whether it's a content creator, a video, or a commentor)...to be the most effective strategy in getting rid of toxicity" (P66). Users also report toxic people to the platform to get them banned and have them "lose the[ir] freedom of speech" (P95, Reddit). Reporting to police is defined as involving law enforcement in toxicity disputes. Participants who suggested this option said to "request the police to take action" (P83, Reddit). Some participants who suggested this option recommended users only involve law enforcement when the toxic content in question is illegal in nature. Irrespective of to whom participants wanted to report toxicity, reporting in general was found to be the easiest, quickest, and most effective strategy for social media users to implement in removing toxicity.

4.2. Platform-based Interventions

Our second major theme was platform-based interventions, which are steps our participants suggested that platforms take to deal with online toxicity. Although some of the specific suggestions are more or less platform-specific (e.g., Reddit already has a downvoting system, while Facebook does not), on the whole, they fall into one of 12 subthemes. There were 45 responses that were categorized under this theme.

4.2.1. Voting Systems Our subtheme of voting systems consists of systems like Reddit's upvote/downvote system that allow users to rank each other and their posts for toxicity. This subtheme had 9 instances. P113 (Facebook) suggested the following as a way to implement a voting system without restricting free speech: "Facebook needs to add thumbs down instead of just having thumbs up for likes. With this, heavily downvoted comments can be hidden or collapsed or put behind a spoiler button without any admin interference."

4.2.2. Warning Systems Connected to this idea is our subtheme warning systems which had 5 instances. These would be systems that platforms put in place to let toxic users know that they have been flagged in some way for their toxic behavior, and that should

the behavior continue, they will be punished. P99 (YouTube) describes one option: "I think if there is a 1 time your [sic.] are out rule. You do something wrong 1 time, the 2nd time, you are banned."

4.2.3. Content Filtering Our next subtheme was content filtering, which consisted of users' suggestions regarding how platforms can add additional options to mute or hide content. One YouTube user (P72) suggested that this should happen automatically during particularly controversial or stressful times (e.g., Trump's proclamation about foreign visas) so as not to breed further toxicity. Several platforms allow their users to filter content based on the users' preferences. For example, on Instagram, users have the option to use a filtering feature called "hide offensive comments" or another called "manual filter" which allows users to list keywords, phrases, and emojis they want hidden from their comments. This subtheme had 2 responses.

4.2.4. Advanced Banning Our theme of advanced banning includes techniques to remove toxic players that go beyond the basic banning options on the platform in question. The most popular type of advanced banning suggested by our participants was an IP ban, which is when a toxic user's IP address is no longer able to access a social media platform. In a similar vein, a Reddit user emphasized "it is imperative to identify users who repeatedly make "alts" [alternate accounts] for ban evasion...Actively preventing these users from making alts would likely do a lot to reduce negativity on social media sites" (P24). Participants want making other accounts to evade a ban more difficult to prevent toxic users from returning to the platform. Participants also identified bots as perpetuating toxicity on social media platforms and mentioned the need for bot removal, stating: "Apparently there are new sites that are good at detecting bots. Twitter needs to use these detectors more and invest in them" (P74, Twitter). There were 8 responses categorized into this subtheme.

4.2.5. Limited Accounts Next is our subtheme limited accounts which had 7 instances. Participants suggested that instead of outright banning, platforms could limit the accounts of toxic users in some way. One participant suggested the following for Reddit to expand upon the existing karma system: "I think there could be a system where people with a certain amount of negative karma have their posting privileges reduced or removed, seeing as they appear to be consistently toxic" (P59).

4.2.6. Financial Punishments Still, other users suggested that offenders should incur financial punishments, another one of our subthemes. This subtheme had 3 responses. For toxic creators, an Instagram user (P7) said that to reduce toxicity, "you have to impact the income stream", while on YouTube, P100 suggested demonetization of videos. Other participants also suggested that toxic users, creators or otherwise, should also be fined.

4.2.7. Targeted Interventions The subtheme targeted interventions is defined as implementing restrictions and guidelines designed to specifically tackle toxicity relating to a particular topic, like racism or sexism, or repeat offenders. For example, P18 (Twitch) wanted Twitch to "quit allowing [the] use of emotes that are different races, the bigotry and racism is fueled by these [sic.]", while a Twitter user (P9) suggested "promoting better education about how offensive tweets can seriously damage a person's mental and physical health." Furthermore, a YouTube user wants the platform to implement "stricter regulations" on popular creators who post content about "body shaming, homophobia, bullying of those with eating disorders, promotion of suicide and bullying of those with mental health issues" (P50). There were 7 instances of this subtheme.

4.2.8. Multi-Step Posting Multi-step posting is our next subtheme which had 3 instances. This was suggested by users who thought that if people had to go through several steps before a post would be approved, there would be less toxicity on social media. Facebook (P144) and YouTube (P6) users both suggested that all comments should need to be reviewed before they are made public, either by the channel/account owner or the company, while P5 (YouTube) affirmed that the verification process on YouTube needs updating.

4.2.9. Reward Systems A user took a different approach and suggested reward systems to support social media users who make a deliberate effort to be kind (P22, Facebook). By doing so, users will be able to focus more on the positive nature of social media rather than the negative. This subtheme only had 1 instance.

4.2.10. Human Moderation, Increased Moderation, and Platform Effectiveness The last three subthemes - human moderation (1 instance categorized into this

subtheme), increased moderation (6 instances), and platform effectiveness (5 instances) - all had to do with platforms either increasing human moderation efforts over algorithmic ones as expressed by "there need to be actual human, independent moderators who work for YouTube, who moderate the videos" (P29, YouTube), increasing moderation of all kinds by hiring more moderators (P131, Twitch), or improving systems that are already in place by making them quicker or more efficient (P58, YouTube).

4.3. Outside Intervention

Our final two major themes did not all include strategies, per se, but did add important nuances to the question of toxicity in social media. The first of these was outside intervention, in which participants suggested that someone completely external to the platform and the users should step in to deal with toxicity. Examples of these interventions are "[improve] the education system in this country" (P52, YouTube), "get a new president" (P82, YouTube), referring to Donald Trump, government-enforced "fines on offenders" (P66, Instagram), and news and media attention to force Reddit admins to take action (P7, Reddit). This theme had 4 responses categorized into it.

4.4. No Intervention

Our last major theme was no intervention, in which participants expressed that toxicity on social media cannot and/or should not be managed. Participants who believed that online toxicity cannot be managed found that toxicity is permanent. According to these users, toxicity is not something social media can ever truly deal with, as expressed by P63 (Twitter), who said: "Twitter is a lost cause on this issue. I just deal with it", and P106 (Twitter), who called Twitter the "toxic cesspool of the internet." Participants who believed online toxicity *should* not be managed defended free speech. These users rejected the premise of the question entirely, claiming that any form of content moderation consisted of a violation of free speech laws in America. P128 (Instagram) expressed it particularly clearly, stating that "'Toxicity' is self-expression and getting rid of it is censorship." Even though participants in these final two major themes did not provide suggestions on how to help reduce or remove toxicity in social media, their opinions are important to take into account when making design decisions regarding social media platforms, as well as when creating policies that involve social media. There are 9 instances of this theme.

5. Contributors to a positive social media experience

For the research question that asked participants what helps to create a positive environment on social media platforms based on their personal experience, we found there to be four major themes: things users can do themselves (8 subthemes), things platforms need to do (2 subthemes), things users and platforms have to work together for (4 subthemes), and the impossibility of a positive environment. The responses categorized into these themes describe the various ways users have either used or seen used in creating positive environments online.

5.1. Things Users Can Do Themselves

5.1.1. Only Interact with People Online that You Know in Real Life A prominent theme was only interacting with people online that you know in real life, as this would remove anonymity. Examples of this theme being practiced were noted throughout the participant responses. A Facebook user stated, "I only interact with friends and family that I know in real life" (P62). Similarly, another user limits their online social circle to only friends and family members. There were 14 instances of this subtheme.

5.1.2. Avoidance of Controversial Topics Our subtheme avoidance of controversial topics is self-explanatory and has the goal of fostering a positive environment by preventing heated arguments. There were 8 instances of this subtheme. One participant suggested users should not "follow anything related to sports, religion, or politics" and rather stick to their hobbies on Instagram (P7). Other participants expressed similar sentiments about politics and religion being controversial topics and the need to avoid their discussion online.

5.1.3. Self-moderation Our next theme, self-moderation, turns its focus to the user and includes various practices the user should employ to create a positive environment. The 6 participants' responses indicate such practices of self-moderation include being aware of what they are posting, holding themselves responsible for their actions (P49, Twitch), being active and listening members of the community (P55, Twitter), practicing patience (P6, YouTube), and either being nice to others or not saying anything at all (P106, YouTube).

5.1.4. Choose Positive Content Responses that indicated participants found that certain types of content create a positive environment and chose to view that specific content were categorized under this subtheme; there were 4 instances of this subtheme. For instance, one Twitter user's response states seeing "positive posts, like cats and dogs, [and] uplifting words" (P109) will create a positive environment while another believes President Trump's tweets make people laugh and smile (P82).

5.1.5. Self-care Our subtheme self-care is defined as users should take care of their emotional and mental health to combat the overwhelming online toxicity and create a positive environment. One participant's response stated a positive environment starts with better health and the absence of stress and other negative factors in a user's life (P68, Twitch). There were 2 instances of this subtheme.

5.1.6. Do Not Use the Platform Our next subtheme is do not use the platform which is defined as not using the platform or having an account to the platform is the only way to have a positive environment. This is demonstrated by P82's response of "Staying off Twitter is the best way. The more active I am on Twitter, the most I feel hopeless about humanity." This subtheme had 2 instances.

5.1.7. Respond with Vengeance This subtheme, respond with vengeance, indicates that some negative people make it difficult to create a positive environment and cannot be handled cordially. There was 1 responses under this subtheme. This user's proposed solution is, "Some people just want to be nasty and angry. They can't be reasoned with, so get rid of them with vengeance" (P133, YouTube).

5.1.8. Education and Communication Our subtheme education and communication has responses that describe user-to-user interactions that explain to the toxic person how to act on social media platforms categorized in it. Participants indicate that if a toxic person is encountered on social media platforms, the rules of the platform or subcommunity should be communicated to the toxic person. This can be done by making a post that will explain the types of content that will be removed and messaging the toxic person directly to have a conversation about the content. This subtheme had 11 instances categorized under it.

5.2. Things Platforms Need To Do

5.2.1. Platform Policy Change Though our subtheme of platform policy change was not prominent with 1 response categorized under it, a participant's response suggests users should be required to pass a test created by the platform before they are able to comment or post content (P82, YouTube). The same participant suggested that people should also be required to sign a contract with the platform that ensures they will behave positively.

5.2.2. Shut Down the Platform One of our more drastic subthemes is shut down the platform. The 2 responses under this theme indicate the only way to create a positive online environment is by shutting down online social media platforms, as they allow for negativity and toxicity. One Twitter user believes "Nothing will help Twitter, it needs to be shut down all together" (P106).

5.3. Things Users and Platforms Have to Work Together For

5.3.1. Better Moderation Practices The most prominent theme was better moderation practices. The 18 responses under this theme indicated that users have various perspectives on what they believe to be good moderation, but they all agree that changing how moderation is currently conducted will create a positive environment. For example, a Reddit user wants "the ability to arrange comments by 'best' and 'controversial'" (P28) while a Twitter user believes a quicker response time from Twitter's moderation team to eliminate toxic content would prevent people from arguing online and harming themselves (P76); though different in their methods, both believe moderation can create positive environments on online platforms.

5.3.2. Responsibility of the Creator/Poster Our next subtheme is responsibility of the creator/poster. There were 7 instances of this subtheme. One response under this theme stated it "starts with the streamer. They cultivate their communities. If they don't spew nonsense themselves and don't enable toxic members, the environment is overwhelming [sic.] positive" (P64). Participants' responses categorized under this

theme indicate that the creator/poster of the content and the community should be responsible for fostering a positive, engaging, and active environment as their content attracts their followers and subscribers.

5.3.3. Limit the Number of Users Another theme is limiting the number of users, as this reduces the number of users interacting with each other. There were 3 instances of this subtheme. This theme applies to the number of people that participate in subcommunities on social media platforms. An example of this is one participant's response that emphasizes a smaller stream on Twitch in which the stream has a connection with the viewers would result in a more positive environment (P101, Twitch). Other Twitch users also suggested that users engage in a "sub only chat" (P18) and the platform should have "fewer users" (P59).

5.3.4. Be Factual The 4 responses under this subtheme of be factual indicate that a positive environment is created online through posts and content that are factual. An Instagram user stated "There's a lot of fake personalities" on the platform and "realness" is needed to combat this (P61). Posters should not post fake or embellished stories about their lives and experiences online, but rather be truthful. If fake content and accounts are found online, they should be removed.

5.4. Positive Environment is Not Possible

Our last major theme is that a positive environment is not possible. There were 3 responses categorized under this theme. One Twitch user states, "toxicity is just part of the appeal of Twitch, so a 'positive' environment includes toxicity" (P148). In a similar sentiment, a Reddit user expressed "I have no control of what others post or comment in the sub. So there's nothing in which I can [do] to create a positive environment" (P37). Responses under this theme indicate that online platforms are going to include toxicity which is a part of the experience and therefore users cannot do anything about it.

6. Discussion

Our data indicates that the ideas users have were not completely novel or outside the box but it's interesting that not all platforms implement or address these in the same way. For example, a reward system for kindness is not something that most social media systems have although it is something that exists in some online games. This suggests that there is an opportunity for platforms to learn from each other and adopt moderation practices that are successful.

Aside from those participants who thought that toxicity was a natural part of life, most participants all voiced a strong opinion toward the company bearing more responsibility. Given that how companies handle negative content may not necessarily be an legal obligation, it was interesting to see users' desire for the company to play a bigger role.

Previous studies have shown that people want law enforcement involvement and stronger laws. In a similar vein, our results show that users believe policing and increase of law enforcement involvement to be an effective strategy. This raises questions about what people think is the appropriate boundary between government/law enforcement involvement and whether our legal system is equipped to process punishment.

While it was somewhat unsurprising that participants suggested both strategies to punish violators as well as reward do-gooders, they also suggested strategies that had nothing to do with engagement with the violator at all and rather focused on the wellbeing of the victim. Building resilience, education, self care, or taking a break from social media, for example, were some of the suggestions along these lines.

6.1. Limitations and Future Directions

Though the current study has significant implications for social media platforms regarding effective toxicity management and content moderation, it does have several limitations. First, we do not consider other forms of online abuse outside of toxicity, such as the existence of fake news or targeted recruitment of social media users to extremist organizations, in the scope of our study. This study was only focused on online toxicity; however, this presents an opportunity for future studies to consider all forms of online abuse. Another limitation to our study is that we do not examine the content moderation preferences of users based on the different platforms they utilize. We also do not ask users about their familiarity with content moderation as learning about their moderation knowledge was not the goal of our study. Examining if users of different platforms have different moderation preferences and if users that have more content moderation knowledge have different preferences are great areas for future study.

7. Conclusion

In this paper, we present a wide range of strategies for removing toxicity from online platforms as well as methods of creating positive social media environments

that social media users believe to be effective. Our results show that participants believe both the platforms and users can and should implement strategies for managing toxicity. Though the strategies presented are not all novel ideas and are often discussed in other studies [6, 8, 29], our results highlight the fact that not all platforms address toxicity management in the same way. While most of our participants have established strategies they utilize to make their online experience less toxic and more positive or expect their platform of choice to manage toxicity, some participants believe toxicity to be an inevitable part of their online experience. This is important to acknowledge as other participants recommend prioritizing health over social media to avoid the negative ramifications of online toxicity. Overall, our results show that most of our participants believe that both platforms and users can and should utilize certain strategies to remove online toxicity and create a more positive environment.

8. Acknowledgement

This research was supported in part by National Science Foundation grants 1928627 and 1841354.

References

[1] Pew Research Center, "Demographics of social media users and adoption in the united states," 2019.

[2] K. Leetaru, "Is social media becoming too toxic?," 2018.

[3] B. Miller, "Countering online toxicity and hate speech," 2019.

[4] D. U. Patton, R. D. Eschmann, C. Elsaesser, and E. Bocanegra, "Sticks, stones and facebook accounts: What violence outreach workers know about social media and urban-based gang violence in chicago," *Computers in human behavior*, vol. 65, pp. 591–600, 2016.

[5] J. Gorner, P. Nickeas, and H. Dardick, "A violent january in chicago: 45 murders and counting," 2016.

[6] M. Duggan, "Online harassment 2017," 2017.

[7] G. Fowler, D. Harwell, E. Dwoskin, and T. Romm, "2018 was the year of online hate. meet the people whose lives it changed.," 2018.

[8] Anti-Defamation League, "Online hate and harassment: The american experience," 2019.

[9] J. Guynn, "If you've been harassed online, you're not alone. more than half of americans say they've experienced hate," 2019.

[10] K. Leetaru, "Is it actually possible to solve online toxicity?," 2019.

[11] A. Arsht and D. Etcovitch, "The human cost of online content moderation," *Harvard Law Review Online, Harvard University, Cambridge, MA, USA. Retrieved from https://jolt. law. harvard. edu/digest/the-human-cost-ofonline-content-moderation.*

[12] S. T. Roberts, *Content Moderation*, pp. 1–4. Cham: Springer International Publishing, 2017.

[13] J. Grimmelmann, "The virtues of moderation," *Yale JL & Tech.*, vol. 17, p. 42, 2015.

[14] S. T. Roberts, *Behind the screen: Content moderation in the shadows of social media*. Yale University Press, 2019.

[15] S. Singh, "How automated tools are used in the content moderation process," 2019.

[16] E. Dwoskin, J. Whalen, and R. Cabato, "Content moderators at youtube, facebook and twitter see the worst of the web — and suffer silently," 2019.

[17] K. Leetaru, "The problem with ai-powered content moderation is incentives not technology," 2019.

[18] J. York and C. McSherry, "Automated moderation must be temporary, transparent and easily appealable," 2020.

[19] M. Perel and N. Elkin-Koren, "Black box tinkering: Beyond disclosure in algorithmic enforcement," *Fla. L. Rev.*, vol. 69, p. 181, 2017.

[20] S. Jhaver, A. Bruckman, and E. Gilbert, "Does transparency in moderation really matter? user behavior after content removal explanations on reddit," *Proceedings of the ACM on Human-Computer Interaction*, vol. 3, no. CSCW, pp. 1–27, 2019.

[21] Santa Clara Principles, "Santa clara principles on transparency and accountability in content moderation.," 2018.

[22] B. Rawlins, "Give the emperor a mirror: Toward developing a stakeholder measurement of organizational transparency," *Journal of Public Relations Research*, vol. 21, no. 1, pp. 71–99, 2008.

[23] T. Gillespie, *Custodians of the Internet: Platforms, content moderation, and the hidden decisions that shape social media*. Yale University Press, 2018.

[24] S. Schoenebeck, O. L. Haimson, and L. Nakamura, "Drawing from justice theories to support targets of online harassment," *new media & society*, p. 1461444820913122, 2020.

[25] A. Glaser, "Want a terrible job? facebook and google may be hiring," *Slate. https://slate.com/technology/2018/01/facebookand-google-are-building-anarmy-of-content-moderatorsfor-2018. html*, 2018.

[26] D. Y. Wohn, "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*, pp. 1–13, 2019.

[27] J. Seering, T. Wang, J. Yoon, and G. Kaufman, "Moderator engagement and community development in the age of algorithms," *New Media & Society*, vol. 21, no. 7, pp. 1417–1443, 2019.

[28] J. LaLoggia, "U.s. public has little confidence in social media companies to determine offensive content," 2019.

[29] J. Cai and D. Y. Wohn, "What are effective strategies of handling harassment on twitch? users' perspectives," in *Conference companion publication of the 2019 on computer supported cooperative work and social computing*, pp. 166–170, 2019.

[30] J. Corbin and A. Strauss, *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage publications, 2014.