Understanding User Preferences for Setting Content Moderation Configurations

Alice Zhang

CRA DREU

Summer 2021 Final Report

Advisor: Amy X. Zhang

September 30, 2021

1 INTRODUCTION

Content moderation in platforms such as Twitter, Twitch, and Facebook plays an important role in regulating harassment, hate-speech, and toxicity [10, 14]. As platforms grow their user bases, content moderation has become crucial to creating spaces for open and free communication.

Generally, content moderation involves reviewing content in online communities to prevent abuse and otherwise harmful behavior [7]. In implementing moderation tools, however, platforms and designers alike must navigate questions of freedom of speech, justice, fairness, and accessibility. Previous work on Twitter Blocklists has found that such tools invoke diverging opinions on both sides. Some users may feel unjustly blocked whereas others expressed concern that the tool is not enough to protect them from harassment [10]. Other work took a framework of justice to uncover user preferences for those violating community guidelines [21]. Around these works, there is much debate on what constitutes justice in the context of content moderation as well as how communities can be kept safe, but not constrict their user bases.

Current Work In this study, we focus on content moderation configurations through a user lens. We explore user preferences for a series of pre-configured content moderation tools and the factors associated with as well as the reasoning behind those choices.

2 RELATED WORK

2.1 Defining Content Moderation

Content moderation encompasses the decisions around whether content posted to platforms should be removed and the implications of such decisions [7]. Discussion around content moderation decisions ask questions of what platforms should have agency over versus users and alternatively what role policy should have [12, 20].

2.2 Harassment in Content Moderation

Twitter blocklists are an example of user-led efforts toward content moderation. This third-party mechanism allows users to block accounts they find harmful. Previous work has found key concerns with blocklists to be user dissatisfaction with the failure of existing moderation tools on Twitter as well as concerns of freedom of expression [10]. Blockparty and Sentropy are third-party moderation mechanisms that work in a similar manner to allow users to filter out words and block or mute users of their choice on Twitter. Squadbox operates with the same focus, but recruits users' friends and family to assist with moderating harmful email content [18].Gobo allows users to configure their news feeds on multiple platforms according to their own preferences [1]. The plethora of tools dedicated to assist users in dealing with harassment show that harassment is a relevant issue to be addressed in content moderation.

2.3 User Control in Social Media

An important area of discussion concerning harassment in content moderation involves the optimization of existing content moderation tools. By comparing user labels of comments from Twitter, Reddit, and 4chan, personalized tuning was found to significantly improve the accuracy of existing mechanisms such as Google Jigsaw's Perspective API [14]. Analysis of how user preferences for support and moderation mechanisms regarding online harassment similarly found that user experiences with harassment and preferences for moderation vastly differ by their individual identities [21]. Further study on the whether the existence of explanations on content moderation configurations affect users' ability to distinguish toxic content found no significant effect [4]. A wide range of concerns exist regarding the lack of transparency on algorithms [3, 17]. These works lead

us to ask the questions: What constitutes user control in content moderation? How can moderation configurations be designed to support users' sense of agency while lessening the burden of labor?

2.4 Techniques for User Control

Analyzing user interactions with and without the presence of control mechanisms, Vaccaro et al. found that users tend to feel more satisfied with results when controls are present regardless of whether they work [23]. Additionally, users' sense of control is also strongly influenced by the type of input mechanism [5]. Such observations illustrate the effect control mechanisms may have on users' sense of control.

2.5 Ethical Implications

Behind the design considerations for content moderation are the ethical implications of such tools. By viewing content moderation through the lens of targets of harassment, previous found concepts and definitions of justice and fairness to vary by users' individual identities. User preferences for how platforms should respond to harassment whether it is greater moderation on a larger scale or empowering users to create their "own space" by controlling what they see varied alongside user perceptions of when harassment was justified if at all [2, 21]. In this context, content moderation may be considered a type of intermediary governance with key questions remaining about its legitimacy. Analysis of online intermediary governance argues for the inclusion of security, dignity, discrimination, and other human rights in addition to the existing focus on freedom of expression and privacy [22].

2.6 Relation to Prior Work

An important distinction must be made with the intersection of our work and previous work. Our investigation into previous work conducted on the effects of design configurations as well as the existence or absence of explanations has lead us to works focused on content moderation as well as recommender systems [4, 8, 13]. However a key distinction between these two sections is that previous work on recommender systems contain an additional product-driven focus whereas our work aims to optimize content moderation configurations for the purpose of user ease of use and protection from toxicity. Previous work mentioned on algorithms may fall into one or both of these categories [15–17].

3 METHODS

We designed a survey which consisted of three parts. The first section asked about participants' experiences with toxicity online and technical aptitude such as programming familiarity. The next section included a guided exercise where participants adjusted a given content moderation configuration to their liking. The last section contained questions about participants' preferences with the design interface, understanding of how the configuration worked, and sense of control and trust, and ethical considerations.

Through this survey, we aim to seek a greater understanding of how decisions about the design of content moderation configurations affect users' perceived sense of agency and burden of labor. Additionally, we will also explore users' perceptions of responsibility in the context of personal content moderation.

3.1 Survey Instrument

We based our survey questions on previous work examining toxic content classification as well as user perceptions of control, fairness, and understandability of algorithmic systems [6, 14–16].

Question on how prior user engagement impacts preferences. If a user spends a lot of time on Twitter, for example, that might contribute to different preferences compared to users who rarely use it. [6] general guide to survey practices [19] asking about prior experience with programming/technical aptitude [16] measured perceived fairness, trust, and emotional response to algorithmic decisions [15] studies with processes similar to ours [11, 13, 14]

Questions in the first section of our survey aimed to understand users' prior knowledge of programming as well as experiences of toxicity and harassment. We included questions on users' experiences with algorithms and programming to examine how users with a greater understanding of how algorithmic configurations would affect their preferences when interacting with such configurations.

Questions in the last section of our survey aimed to understand users' experience interacting with their given content moderation configuration. We asked users about their understanding of the configuration and their likelihood of using the specific tool on certain platforms. Platforms were asked about included: Twitter, Twitch, and Instagram. Additionally, we sought out users' perceptions of control and trust in the configuration tool. We constructed a series of [number] hypotheses on users; perceptions of trust: [hypothesis]. Based on these hypotheses, we asked users to rate their agreement or disagreement on a seven-point Likert scale. We also utilized the NASA TLX to ask users about the burden of labor. We decided to use the questions about [topics] in the NASA TLX [9].

Finally, we asked users about their perceptions of responsibility in relation to personal content moderation. We investigated where users perceived the responsibility of content moderation should fall as well as what categories of toxicity users would want control over versus which categories they would prefer for platforms to moderate. Finally, we also asked users whether the existence of a personal content moderation tool on a platform would make them more likely to speak out.

We designed our questionnaire to avoid common biases as outlined by Muller et al [19]. For instance, participant responses were anonymous to minimize social desirability bias. To minimize question order bias, we place open-ended and complex questions toward the end of our survey. To minimize satisficing varied the structure of our questions, placing questions in alternating order by type. We also tracked the time users spent on the survey as well as the time users spent on the exercise provided. Finally, we excluded broad, learning and double-order questions.

We also constructed all our agree/disagree questions to be on bipolar constructs, which range from an extreme negative to an extreme positive rather than unipolar constructs, which range from zero to an extreme amount. Bipolar constructs also involve seven-point Likert scales whereas unipolar constructs are best measured with five-point rating scales.

3.2 Guided Exercise

The second section of our survey involved a guided exercise. Users were placed in one of four conditions: word filter, binary, intensity, or probability. Each condition involved a different content moderation configuration. The Intensity and probability conditions both featured a slider option and simulated an algorithmic moderation tool.

Users were first presented with a condition interface and a set of pre-selected comments. To simulate natural interaction on platforms, users could go back and forth between the "feed" of pre-selected comments and the "settings" where the configuration was. Users were given [time] to adjust until they were satisfied with their feed. As mentioned previously, the time users spend adjusting their feed was tracked.

Users were then presented with a different set of comments filtered based on their adjustment of the configuration before. A series of follow-up questions were asked in the last section of the survey on users' experiences with the configuration.

3.3 Recruitment and Data Collection

Our sample consists of participants from Amazon Mechanical Turk. We decided to pay participants $2.5 for filling out what we estimated to be a 10-minute survey. Qualifications for participants required that they be at least 18 years of age, in the U.S., and English speakers. We also provided participants with the option to decline to answer demographic questions.

3.4 Consent

Participants were required to sign a content form before participating in the survey. The form validated their age as well as qualifications. Before participants began the survey, we included a warning describing the type of toxic content they may be exposed to:

Risks related to this research include exposure to potentially toxic comments which may result in participants feeling targeted, hurt, and/or recalling negative personal experiences with toxic comments online. Participants have the right to exit the survey at any time if they feel the need do.

Participants were also given the option to exit the survey at any time should they feel the need to.

3.5 Variables

We measured five dependent variables in our study which include: explanation effectiveness, stickiness, sense of control, and ethical considerations.

4 NEXT STEPS

The following report details the findings and preparations for the completion of our study in the following academic year. We will obtain results and perform data analysis this upcoming year building off of our work this summer.

REFERENCES

[1] Rahul Bhargava, Anna Chung, Neil S Gaikwad, Alexis Hope, Dennis Jen, Jasmin Rubinovitz, Belén Saldías-Fuentes, and Ethan Zuckerman. 2019. Gobo: A system for exploring user control of invisible algorithms in social media. In *Conference Companion Publication of the 2019 on Computer Supported Cooperative Work and Social Computing*. 151–155.

[2] Lindsay Blackwell, Jill P Dimond, Sarita Schoenebeck, and Cliff Lampe. 2017. Classification and Its Consequences for Online Harassment: Design Insights from HeartMob. *PACMHCI* 1, CSCW (2017), 24–1.

[3] Jenna Burrell, Zoe Kahn, Anne Jonas, and Daniel Griffin. 2019. When users control the algorithms: values expressed in practices on twitter. *Proceedings of the ACM on Human-Computer Interaction* 3, CSCW (2019), 1–20.

[4] Samuel Carton, Qiaozhu Mei, and Paul Resnick. 2020. Feature-Based Explanations Don't Help People Detect Misclassifications of Online Toxicity. In *Proceedings of the International AAAI Conference on Web and Social Media*, Vol. 14. 95–106.

[5] David Coyle, James Moore, Per Ola Kristensson, Paul Fletcher, and Alan Blackwell. 2012. I did that! Measuring users' experience of agency in their own actions. In *Proceedings of the SIGCHI conference on human factors in computing systems*. 2025–2034.

[6] Motahhare Eslami, Kristen Vaccaro, Min Kyung Lee, Amit Elazari Bar On, Eric Gilbert, and Karrie Karahalios. 2019. User attitudes towards algorithmic opacity and transparency in online reviewing platforms. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*. 1–14.

[7] James Grimmelmann. 2015. The virtues of moderation. *Yale JL & Tech.* 17 (2015), 42.

[8] Jaron Harambam, Dimitrios Bountouridis, Mykola Makhortykh, and Joris Van Hoboken. 2019. Designing for the better by taking users into account: A qualitative evaluation of user control mechanisms in (news) recommender systems. In *Proceedings of the 13th ACM Conference on Recommender Systems*. 69–77.

[9] Sandra G Hart and Lowell E Staveland. 1988. Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology*. Vol. 52. Elsevier, 139–183.

[10] Shagun Jhaver, Sucheta Ghoshal, Amy Bruckman, and Eric Gilbert. 2018. Online harassment and content moderation: The case of blocklists. *ACM Transactions on Computer-Human Interaction (TOCHI)* 25, 2 (2018), 1–33.

[11] Yucheng Jin, Nava Tintarev, and Katrien Verbert. 2018. Effects of personal characteristics on music recommender systems with different levels of controllability. In *Proceedings of the 12th ACM Conference on Recommender Systems*. 13–21.

[12] Sara Kiesler, Robert Kraut, and Paul Resnick. 2012. Regulating behavior in online communities. *Building Successful Online Communities: Evidence-Based Social Design* (2012).

[13] Bart P Knijnenburg, Niels JM Reijmer, and Martijn C Willemsen. 2011. Each to his own: how different users call for different interaction methods in recommender systems. In *Proceedings of the fifth ACM conference on Recommender systems*. 141–148.

[14] Deepak Kumar, Patrick Gage Kelley, Sunny Consolvo, Joshua Mason, Elie Bursztein, Zakir Durumeric, Kurt Thomas, and Michael Bailey. 2021. Designing Toxic Content Classification for a Diversity of Perspectives. *arXiv preprint arXiv:2106.04511* (2021).

[15] Min Kyung Lee. 2018. Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society* 5, 1 (2018), 2053951718756684.

[16] Min Kyung Lee and Su Baykal. 2017. Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division. In *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*. 1035–1048.

[17] Min Kyung Lee, Daniel Kusbit, Evan Metsky, and Laura Dabbish. 2015. Working with machines: The impact of algorithmic and data-driven management on human workers. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*. 1603–1612.

[18] Kaitlin Mahar, Amy X Zhang, and David Karger. 2018. Squadbox: A tool to combat email harassment using friendsourced moderation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.

[19] Hendrik Müller, Aaron Sedley, and Elizabeth Ferrall-Nunge. 2014. Survey research in HCI. In *Ways of Knowing in HCI*. Springer, 229–266.

[20] John Samples. 2019. Why the government should not regulate content moderation of social media. *Cato Institute Policy Analysis* 865 (2019).

[21] Sarita Schoenebeck, Oliver L Haimson, and Lisa Nakamura. 2021. Drawing from justice theories to support targets of online harassment. *new media & society* 23, 5 (2021), 1278–1300.

[22] Nicolas Suzor, Tess Van Geelen, and Sarah Myers West. 2018. Evaluating the legitimacy of platform governance: A review of research and a shared research agenda. *International Communication Gazette* 80, 4 (2018), 385–400.

[23] Kristen Vaccaro, Dylan Huang, Motahhare Eslami, Christian Sandvig, Kevin Hamilton, and Karrie Karahalios. 2018. The illusion of control: Placebo effects of control settings. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*. 1–13.