How does providing users with the choice to avoid toxic political content impact their experience on social media?

Keywords: User choice, Personal content moderation, User experience, Algorithmic choice, Social media

Extended Abstract

There is mounting concern about the prevalence of hateful, uncivil, or toxic content on social media. Many call for social media platforms to provide users with the choice to avoid such content within their newsfeeds in order to decrease polarization, the spread of misinformation, and other negative behaviors online. Yet scholars have not yet examined whether such interventions will have such positive impacts, and whether they may create unintended consequences. We present a series of experiments designed to examine how providing social media users with the choice to avoid content shapes their opinions about political content and their overall satisfaction with platforms. Together, our experiments demonstrate that social media users like having the option to avoid toxic content, but—paradoxically— those who choose to avoid such content express more negative evaluations, not less. These findings have important implications for the study of political polarization, social media, and computational social science— as well as technology companies, politicians, and other practitioners who are contemplating how to regulate the social media industry.

Most social media companies have a form of content moderation (e.g., banning harmful users, removing graphic or violent content, adding clarifying comments under potentially misleading posts etc.). However, the public's growing weariness over the ethics and potential harms of social media have pushed platforms to try to find alternative solutions to content moderation. One such solution is to give users more control and agency over the types of content that they consume. Recently, for example, Intel introduced Bleep, an AI tool that helps gamers avoid toxic content and hate speech in online voice chats. The tool consists of toggles that users can move to dictate how much they would like to see of various forms of hate speech and toxic content. Intel's announcement brought in a wide range of reactions (Zhang, forthcoming). Some felt like the tool puts too much onus on the user to dictate how much hate speech they were willing to face, removing responsibility from companies, while others welcomed the tool as it would provide them with the means necessary to have a safe space online.

While these developments may be a step in the right direction, very little is known about the impacts of such changes on users and their online experiences. Research on user choice and personal content moderation is still in its very early stages, so it is all the more imperative that researchers investigate the potential consequences (Jhaver and Zhang 2023; Einwiller and Kim 2020; Riedl et al. 2021). With this study, we test two main hypotheses.

H1: Giving users the perception they are controlling their user-experience may lead to more positive evaluations of the platform.

H2: Giving users the perception that they are controlling whether they see toxic political content may lead to more positive evaluations of polarizing content.

There is reason to suggest that users may welcome the increased agency. A recent representative survey of US internet users, for example, showed that 52.4% prefer to moderate the content they consume online over relying on platforms to do the moderating for them (Jhaver and Zhang 2023). Furthermore, past research has shown that users appreciate being given the ability to control aspects of recommender algorithms (e.g., choosing favorite genres when browsing films online) (Knijnenburg et al., 2012; Dooms et al. 2014). It also increases their trust in the platforms (McNee et al 2003; Xiao and Benbasat 2007).

We recruited a total of 2484 participants to take part in two survey-experiments where they were asked to provide feedback on a (fake) new social media platform and its algorithm. Participants were asked to evaluate various social media posts from this platform. These posts consisted of eight political and four non-political posts. The treated half of our participants were given the option of evaluating posts that were specially selected by a new algorithm—one that is designed to show fewer politically toxic posts. The other half, in our control group, were not given this option. In actuality, all participants saw the exact same posts. We found that regardless of the type of content they saw, those who explicitly chose to opt-in to viewing less toxic content tended to rate political and neutral posts more negatively, even though these same participants reported higher satisfaction with this new platform.

We hypothesized that these differences may have been due to "broken promises." That is, since we told treated participants that the algorithm would prioritize positive content, participants may have been upset to see a sprinkling of negative posts or to see posts that were not completely positive. Giving people choice may have created high expectations for the technology, which may have been polarizing when those expectations were unmet. To investigate the potential "broken promises" effect, we ran a follow-up experiment. This study was identical to the original, however, we removed all negative political posts from the evaluation portion. We found that treated participants who opted-in to viewing less toxic posts *still* evaluated political and neutral posts more negatively than participants in control, while also reporting higher satisfaction with the new platform.

References

Dooms, S., Pessemier, T. D., and Martens, L. (2014), "Improving IMDb Movie Recommendations with Interactive Settings and Filters." RecSys Posters.

Einwiller SA and Kim S (2020) How Online Content Providers Moderate User-Generated Content to Prevent Harmful Online Communication: An Analysis of Policies and Their Implementation. Policy & Internet 12(2): 184-206.

Jhaver, S., and Zhang, A. (2023), "Do Users Want Platform Moderation or Individual Control? Examining the Role of Third-Person Effects and Free Speech Support in Shaping Moderation Preferences," arXiv. http://arxiv.org/abs/2301.02208

Knijnenburg, B. P., Bostandjiev, S., O'Donovan, J., and Kobsa, A. (2012), "Inspectability and control in social recommenders," in Proceedings of the sixth ACM conference on Recommender systems - RecSys '12, Dublin, Ireland: ACM Press, p. 43.

McNee, S. M., Lam, S. K., Konstan, J. A., and Riedl, J. (2003), "Interfaces for Eliciting New User Preferences in Recommender Systems," in User Modeling 2003, Lecture Notes in Computer Science, pp. 178–187. https://doi.org/10.1007/3-540-44963-9_24.

Riedl MJ, Naab TK, Masullo GM, et al. (2021) Who is responsible for interventions against problematic comments? Comparing user attitudes in Germany and the United States. Policy & Internet.

Xiao, B., and Benbasat, I. (2007), "E-Commerce Product Recommendation Agents: Use, Characteristics, and Impact," MIS Quarterly. https://doi.org/10.2307/25148784.

Zhang, Amy (forthcoming) Understanding User Preferences for Personal Content Moderation Controls.