Experiments offering social media users the choice to avoid toxic political content

Fatima Alqabandi[1,2], **Graham Tierney**[2], **Christopher Bail**[*1,3,4], **Sunshine Hillygus**[3,4], **and Alexander Volfovsky**[2,5]

[1]Department of Sociology, Duke University, Durham, NC, USA
[2]Department of Statistical Science, Duke University, Durham, NC, USA
[3]Department of Political Science, Duke University, Durham, NC, USA
[4]Sanford School of Public Policy, Duke University, Durham, NC, USA
[5]Department of Computer Science, Duke University, Durham, NC, USA
[*]christopher.bail@duke.edu

ABSTRACT

Amid growing concerns that social media algorithms amplify toxic content, platforms are experimenting with giving users more control over their content exposure. This study examines how promises of algorithmic control influence user satisfaction and content evaluation. Offering users the option to avoid toxic political content increases platform satisfaction. However, those who chose this filtering option rated identical posts as more hostile than those without choice. Respondents exposed only to positive posts also rated them as more hostile than those who saw both positive and negative ones. These findings challenge assumptions about user autonomy, showing how promises of control shape expectations and interpretations. Users offered algorithmic control may hold content to stricter standards or shift from passive consumers to active evaluators, attempting to "train" the algorithm. This suggests content curation reflects complex interactions between users and algorithms, where increased autonomy can paradoxically lead to more critical evaluations even as it improves satisfaction.

Introduction

Today's digital public spheres operate under the influence of complex algorithmic systems. These algorithms, designed to maximize engagement, often amplify divisive or hostile content, fueling political polarization and widening ideological divides (Auxier, 2020; Rathje et al., 2021; Brady et al., 2021; Bode, 2016; Cheng et al., 2017; Rathje et al., 2024; Harris et al., 2023). By prioritizing content that triggers strong emotional reactions—like outrage or hostility– algorithms create perverse incentives for users to adopt toxic behaviors in a bid for virality. Negativity spreads quickly in this environment, fueling cycles of moral contagion that amplify hostility and polarization across platforms (Brady et al., 2021; Rathje et al., 2021). These systems don't just reflect existing divisions– they actively deepen them, contributing to the widespread perception that social media is harming society by fostering animosity and polarization (Auxier, 2020).

Despite these challenges, users are not passive consumers; they engage with content in active, nuanced ways. Recent developments, introduced by social media companies in response to a growing awareness of the negative impacts of their platforms, suggest a potential shift towards user autonomy: some platforms now allow users to exert more direct control over the algorithms that influence what content they see. For example, Bluesky, a decentralized platform very similar to Twitter, lets users directly customize their algorithms and curate their feeds (Ahmed, 2024; Cole, 2024). TikTok recently rolled out a "Manage Topics" feature, giving users the ability to adjust how much of specific content categories show up in their "For You" feed, making it easier to tailor their experience (Sato, 2024), while Meta and X (formerly Twitter) have increased transparency in how its recommendation systems work (Meta, 2020). This shift follows years of users developing their own "folk theories" about how algorithms work and attempting to exercise agency within platform constraints (Bucher, 2017; Eslami et al., 2016).

These developments reflect broader tensions in how platforms navigate user agency and algorithmic control. Social media companies must balance their reliance on engagement-based algorithms that drive advertising revenue with growing demands for user autonomy. While increasing user control might reduce behaviors that boost profits (Gillespie, 2010), it could also encourage users to remain on the platform by mitigating their exposure to negativity. This raises an important question: does giving users more control over their online experiences ultimately reduce engagement, or could it enhance user retention by creating healthier environments?

While platforms now increasingly offer users interface-level controls (e.g. Instagram's "Content Preferences" options, TikTok's "not interested" button), these features often operate within systems still fundamentally optimized for engagement metrics rather than user preferences (Van Dijck et al., 2018). This creates a complex dynamic where users must actively develop and test their own theories about how these algorithmic systems work in order to exercise meaningful agency within platform constraints (Eslami et al., 2016; Bucher, 2017). For example, users might experiment with which posts they interact with to influence future recommendations - a strategy that platforms are increasingly acknowledging and formalizing. Instagram, for instance, published a blog confirming that user actions like spending time on posts, commenting, liking, and resharing directly influence content prioritization in their feeds (Con, 2022). This shift from implicit user theories to explicit platform confirmation reflects the growing recognition of users as active participants in algorithmic systems.

Recent research suggests there is broad public support for giving users more control over social media algorithms, but the actual impacts of implementing such controls remain largely untested (Rathje et al., 2024; Jhaver and Zhang, 2023). In a nationally representative study, Rathje et al. (2024) found that nearly 87% of Americans believe users should have "more control over how social media algorithms work." This aligns with a broader pattern of users actively seeking ways to exercise agency over their social media experiences, whether through official platform controls or by developing their own tactics for algorithmic engagement (Gerrard, 2018; Bucher, 2017; Eslami et al., 2016). However, as Rathje and colleagues (2024) note, these solutions need rigorous testing, as they may have unintended consequences—for example, such reforms may encourage people to self-select into echo chambers or spiral down "conspiratorial rabbit holes" (p. 790).

The relationship between user choice and broader issues like echo chambers and algorithmic bias remains understudied. Understanding how offering choice impacts individual user experience is an important first step in addressing these challenges. Our study helps address this critical gap by experimentally examining how offering users the choice to avoid toxic political content affects their experiences and perceptions.

Our study explores how the mere promise of algorithmic control influences both users' satisfaction with a hypothetical social media platform called ConversationCircle, and their evaluation of its content. Through two experiments, where we look at the effect of offering users the option to filter out toxic political content (defined here as posts containing hostile or derisive language aimed at political groups), we uncover a striking paradox: giving users the ability to filter out toxic content increased their satisfaction with our platform, however, for those who actually chose to use the filter, this choice increased their perceptions of hostility in the content they viewed— even though all respondents saw the same content regardless of their treatment condition.

In our second experiment, where we excluded negative content entirely, we found that users who were offered the choice and accepted it, were no less likely to rate the content they viewed as hostile. In fact, when we compared the hostility ratings of all non-negative posts from respondents in our first study to those in our second, we found that respondents in the latter (who saw no negative content) rated all positive posts as more hostile than those in the first. Combined, our findings suggest that giving users more autonomy does not automatically improve their online experiences. Instead, it appears that offering users the choice to filter out content can elevate their expectations in ways that increase, rather than decrease, perceived negativity. These results contribute to the growing literature on algorithms and agency, showing how user autonomy can reshape not only platform experiences but also expectations about content. Building upon emerging research on the impact of algorithmic systems, we add new insights into how perceived control shapes user perceptions and engagement behaviors.

Background

Social Media Platforms and Algorithmic Systems

Social media platforms like Facebook, X (formerly Twitter), and TikTok have transformed Habermas's concept of public spheres—spaces for rational, collective discussion—into digital forums accessible to billions. This democratization of public discourse has helped grassroots movements like the Arab Spring and the #MeToo movement gain traction and widespread support (Papacharissi, 2008; Reynolds, 2007). However, these platforms have, at times, fallen short of Habermas's ideal. Instead of fostering rational and healthy discourse, critics argue that these platforms often reflect the influence of algorithms which tend to prioritize sensational content for their engagement– contributing to the spread of misinformation, divisive content, and political polarization (Rathje et al., 2021; Brady et al., 2021; Cheng et al., 2017; Auxier, 2020).

These platforms shape public discourse not just through content moderation, but through what Seaver (2019) calls “algorithmic systems.” These systems are not “standalone little boxes, but massive, networked ones with hundreds of hands reaching into them, tweaking and tuning, swapping out parts and experimenting with new arrangements” (Seaver, 2019, p.419). Users are not passive recipients of algorithmic decisions but active participants who develop their own “algorithmic imaginaries”– ways of thinking about and experiencing algorithms in everyday life (Bucher, 2017). For example, when TikTok users deliberately watch certain videos longer to “train” their “For You” page, or when Instagram users interact with specific posts to signal their preferences to the algorithm, they are acting on their imaginaries about how these systems work.

When platforms promise users control over their experiences, they must balance this promised autonomy against their need for more engagement and subsequent algorithmic development (Van Dijck et al., 2018). However, viewing these dynamics as simply algorithmic control versus user agency oversimplifies the complex socio-technical nature of these systems. Even platform engineers cannot fully predict nor explain how these systems will behave, as they are "works of collective authorship, made, maintained, and revised by many people with different goals at different times" (Seaver, 2019, p.418). Users, through their interactions and interpretations, become part of these systems, creating what Bucher (2017) describes as a recursive relationship between user behavior and algorithmic operations. For example, when users purposely avoid specific types of content, the algorithm learns from this behavior, which in turn shapes what content users see and how they interact with the platform in the future.

Platform Control and User Agency

Platforms operationalize user agency through various features that promise direct influence over algorithmic systems (for example, Instagram’s "Not Interested" button allows users to theoretically reduce similar content in their feeds). When TikTok allows users to influence their "For You" feed, for example, it frames content selection as user-driven, but it continues to preserve the underlying engagement-based algorithm. This tension between promised and actual control shapes how users understand and navigate platform spaces (Bucher, 2017; Eslami et al., 2015).

The recent platform developments reflect the growing pressure to give users more control. Following whistleblower testimony about social media's negative impacts, a pressing dialogue has emerged among industry leaders, scholars, and policy makers about reform (Persily and Tucker, 2020; Beknazar-Yuzbashev et al., 2022; C-Span, 2021; Allyn, 2021). Acknowledging tensions between content moderation and free speech, some advocate for increased user control over content-shaping algorithms (Jhaver et al., 2023; Jhaver and Zhang, 2023). While most platforms allow basic controls like blocking unwanted interactions, several companies have introduced more sophisticated features. Intel recently offered video game users AI-powered tools to filter hate speech in live audio chats. Rather than relying on opaque algorithms, Bluesky allows users to directly customize their feeds, offering "an open marketplace of feeds" that users can subscribe to, pin, and sort based on their preferences. For example, users can follow feeds dedicated to live events like sports games or TV show premieres, or create feeds that prioritize posts from their mutual connections instead of "viral posts across the whole network" (Bluesky, 2023). These changes mark a significant shift from the engagement-driven, opaque systems of more established platforms, signaling a broader shift toward user control.

The Promise and Perils of User Control

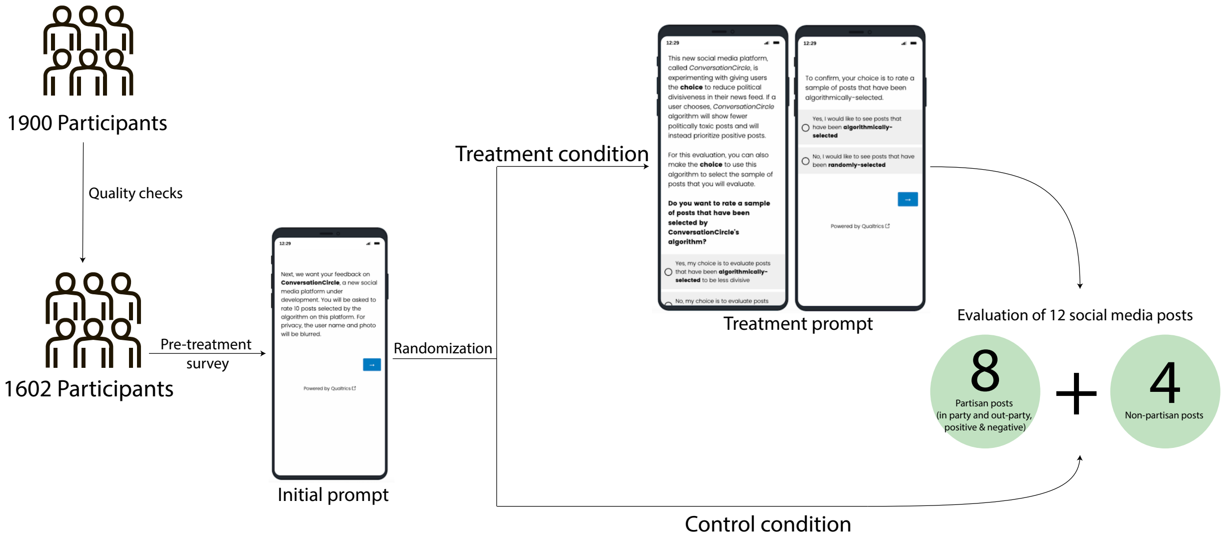
Social science research suggests that providing people with choice (or even the illusion of autonomy) will produce a range of positive outcomes. These include increased motivation and subjective well-being, improved experience, as well as enhanced performance at small tasks (Langer and Rodin, 1976). For instance, Langer (1975) showed that lottery respondents who were given the choice to select their own ticket (as opposed to being given one by the researchers) reported higher confidence in winning. Similarly, Dember et al. (1992) found that giving respondents the choice between "hard" and "easy" versions of a task led to increased commitment and persistence. This phenomenon extends to online environments as well. People appear to appreciate the ability to influence recommender algorithms that might help them explore films, for example (Knijnenburg et al., 2012; Dooms et al., 2014). Research also suggests that giving users more control fosters greater trust in technology platforms (McNee et al., 2003; Xiao and Benbasat, 2007).

However, there are some reasons to think that the reaction of social media users could be less straightforward. Consider the challenge of filtering “toxic” content: platforms must first define what constitutes toxicity, determine acceptable thresholds, and translate these subjective judgments into algorithmic rules. As Seaver (2019) notes, such seemingly simple features require complex cultural interpretations: "How is it defined? What experiences do [developers] draw on as they attempt to formalize it? How might these formalizations differ if we started from different assumptions and experiences?" (p.418). Prior research finds that individuals become frustrated when promoted social media content does not meet their expectations (Bailey et al., 2016; Nanz and Matthes, 2022). Given the subjective nature of what constitutes “toxic” content and how users develop their own interpretations of algorithmic systems (Bucher, 2017), then, there is a possibility that offering users the choice to avoid some content could backfire if users have different expectations about what constitutes toxic content or how much control they should have.

Research Design

We conducted two survey preregistered experiments to evaluate the effect of offering users the choice to avoid toxic political content. As Figure 1 shows, we recruited 1900 Americans over age 18 who identified with either the Democratic or Republican party via CloudResearch, a participant recruitment platform. We asked these respondents to complete a survey covering various topics in politics and social media (e.g., political ideology, their opinions about online polarization, divisiveness, their support of using algorithms as a tool to moderate online content, and other measures that describe their social media experiences). Please see Supplementary Information 8 for the full survey instrument.

Next, respondents were asked to evaluate what they were told was a new social media platform in development called ConversationCircle. We presented respondents with twelve fabricated posts meant to simulate typical platform content. Half the respondents were randomized into the control condition where they did not have a choice over the type of posts they consumed (far left panel of Figure 2). The other half were assigned to the treatment condition, where they were told that ConversationCircle "is experimenting with giving users the choice to reduce political divisiveness in their news feed." Treated users were told this option would reduce the number of politically toxic posts in their timelines and prioritize positive posts instead (right two panels of Figure 2). All respondents viewed *the same posts*, regardless of their experimental condition or choice. This design allows us to evaluate how the perception of choice shapes assessments of political content and the platform overall. This design also allows us to better understand the effect of offering the choice to avoid toxic political content on those who actively choose to do so.



Next, we want your feedback on **ConversationCircle**, a new social media platform under development. You will be asked to rate 10 posts selected by the algorithm on this platform. For privacy, the user name and photo will be blurred.

This new social media platform, called *ConversationCircle*, is experimenting with giving users the **choice** to reduce political divisiveness in their news feed. If a user chooses, *ConversationCircle* algorithm will show fewer politically toxic posts and will instead prioritize positive posts.

For this evaluation, you can also make the **choice** to use this algorithm to select the sample of posts that you will evaluate.

Do you want to rate a sample of posts that have been selected by ConversationCircle's algorithm?

○ Yes, my choice is to evaluate posts that have been **algorithmically-selected** to be less divisive

○ No, my choice is to evaluate posts

To confirm, your choice is to rate a sample of posts that have been algorithmically-selected.

○ Yes, I would like to see posts that have been **algorithmically-selected**

○ No, I would like to see posts that have been **randomly-selected**

Figure 1. We recruited 1900 Democrats and Republicans over age 18 in the United States via Cloudresearch. After eliminating those who failed two or more quality checks, we were left with 1602 respondents. Respondents were first asked a series of questions about their opinions on social media and politics. They were then shown a prompt (the "Initial prompt") that explained that we want their "feedback on ConversationCircle, a new social media platform under development." They were then told they would "be asked to rate 12 posts selected by the algorithm on this platform." Respondents in the treatment condition were given a choice over the types of posts they were to evaluate (the "Treatment prompt"). They could choose between evaluating posts that have been algorithmically-selected to show fewer politically toxic posts or posts that have been randomly-selected. All respondents were then taken to the evaluation portion of the experiment, where they rated the same posts, regardless of experimental condition and choice.

All respondents viewed a total of twelve posts (for a full list of these posts, please see Supplementary Information

4). The posts were identical for all respondents with the exception that the referenced party was dependent on the respondents' reported partisanship.[1] Eight of these posts were political in nature, and four were non-partisan.[2] Of the eight political posts, half were about the respondent's own party and the other half were about the other party (there were two posts each that consisted of positive in-party, positive out-party, negative in-party, and negative out-party in content).

To measure user satisfaction with the platform, we asked people how likely or unlikely they would be to use the platform and whether they would recommend it to friends or family using a seven-point ordinal scale that ranged from "Extremely unlikely" (-3) to "Extremely likely" (3). To form a composite measure of the respondents' *overall* satisfaction with the platform, we also report the maximum of the two scales to capture a strong positive response to either of these dimensions— essentially assessing whether a respondents is likely to use the platform themselves or would recommend it to others. Our second key outcome variable is the mean hostility score across all posts (we also include mean hostility ratings for the different posts types: political, non-partisan, as well as in- and out-partisan posts). To form these hostility ratings, we asked respondents to evaluate how hostile each post was using a five-point ordinal scale that ranged from Not at all Hostile (0) to Extremely Hostile (4).

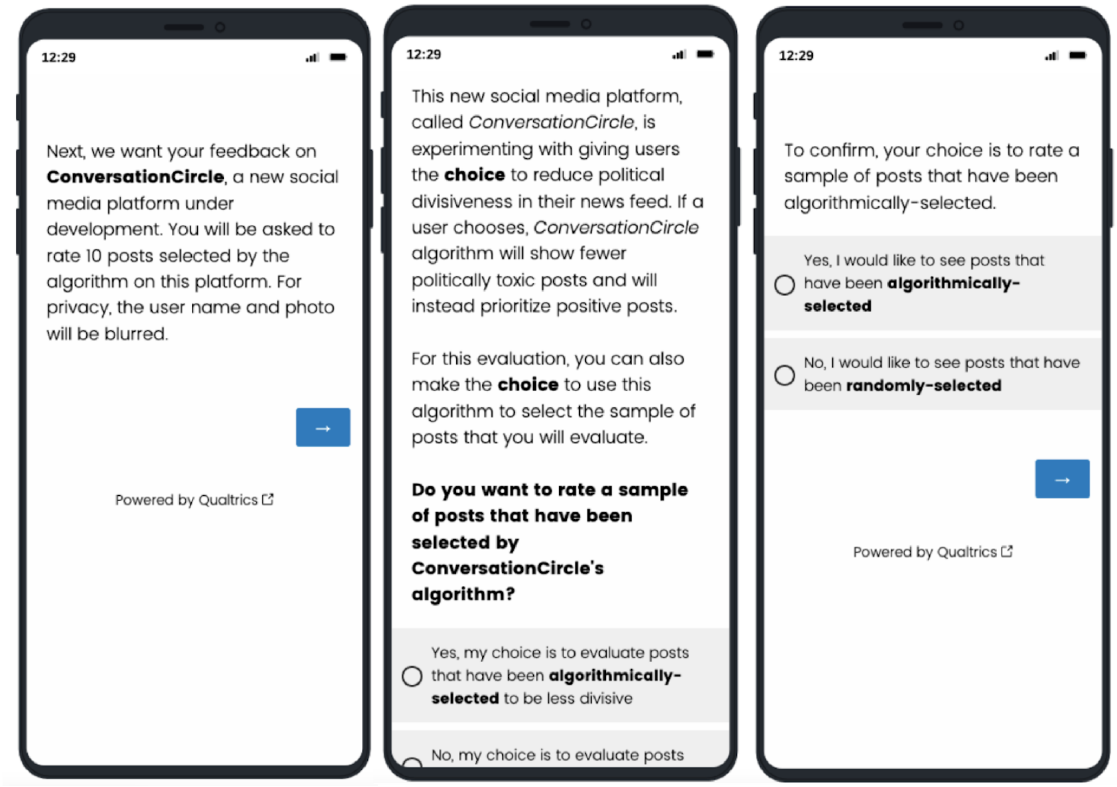
Strengthening our treatment

We took some steps to ensure that our treated respondents were being effectively treated—that is, we made sure the "choice" aspect of the prompts was salient amongst those in the treatment condition. After treated respondents made their choice over whether to see algorithmically-selected posts, they were asked to confirm their choice (Figure 2):

> To confirm, your choice is to rate a sample of posts that have been algorithmically [randomly]-selected.
>
> - Yes, I would like to see posts that have been algorithmically [randomly] selected.

[1] For example, for an in-party negative post, Democrats saw a post that said "If you vote Democrat, you're an idiot. I'm sorry, I do not make the rules." Republicans saw the same post, with "Republican" mentioned instead of "Democrat."

[2] The non-partisan posts consisted of non-political and non-partisan posts (i.e., they didn't say anything that favored one party over another)


12:29

Next, we want your feedback on **ConversationCircle**, a new social media platform under development. You will be asked to rate 10 posts selected by the algorithm on this platform. For privacy, the user name and photo will be blurred.

→

Powered by Qualtrics

12:29

This new social media platform, called *ConversationCircle*, is experimenting with giving users the **choice** to reduce political divisiveness in their news feed. If a user chooses, *ConversationCircle* algorithm will show fewer politically toxic posts and will instead prioritize positive posts.

For this evaluation, you can also make the **choice** to use this algorithm to select the sample of posts that you will evaluate.

Do you want to rate a sample of posts that have been selected by ConversationCircle's algorithm?

○ Yes, my choice is to evaluate posts that have been **algorithmically-selected** to be less divisive

○ No, my choice is to evaluate posts

12:29

To confirm, your choice is to rate a sample of posts that have been algorithmically-selected.

○ Yes, I would like to see posts that have been **algorithmically-selected**

○ No, I would like to see posts that have been **randomly-selected**

→

Powered by Qualtrics

Figure 2. Respondents in control were only shown the first prompt on the left. Those in treatment were shown all three prompts. We gave these treated users the "choice" to see posts selected by this algorithm. Once they answered, we then confirmed their choice to ensure that these users were effectively being treated

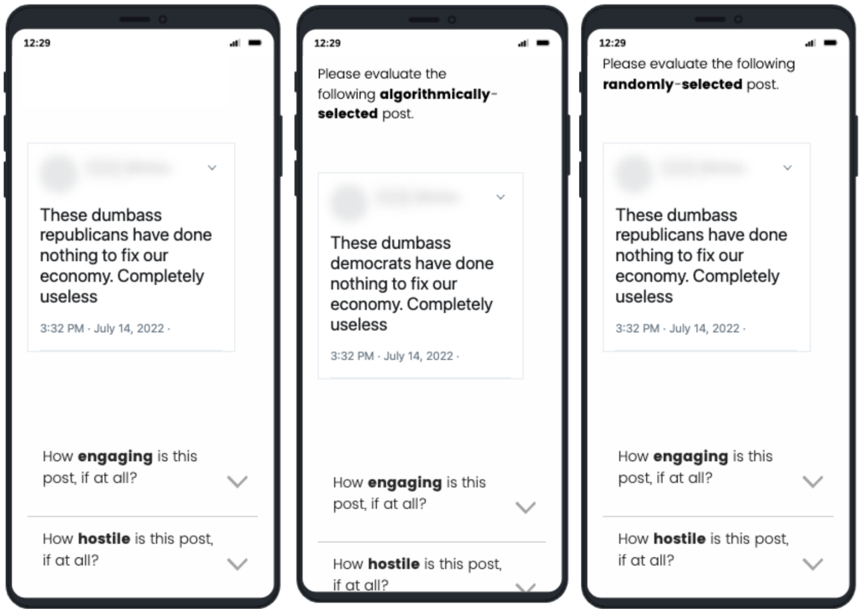
- No, I would like to see posts that have been randomly [algorithmically] selected.

Additionally, in the evaluation portion of the survey (Figure 3), treated respondents were shown a header above each post that stated:

> Please evaluate the following algorithmically-selected [randomly-selected] post.

At the end of the survey, all respondents were also asked to think back to the evaluation portion of the survey and to recall if they were ever given a choice over the types of posts. If they answered yes, they were asked if they remembered which choice they made. We found that 95% of our respondents answered these questions correctly.

The study was approved by our university's Institutional Review Board, and respondents provided informed consent to participate. They were debriefed about the nature of the study and our experimental manipulation after the conclusion of the experiment. All methods were performed in accordance with the relevant guidelines and regulations.



12:29

These dumbass republicans have done nothing to fix our economy. Completely useless

3:32 PM · July 14, 2022 ·

How **engaging** is this post, if at all?

How **hostile** is this post, if at all?

12:29

Please evaluate the following **algorithmically-selected** post.

These dumbass democrats have done nothing to fix our economy. Completely useless

3:32 PM · July 14, 2022 ·

How **engaging** is this post, if at all?

How **hostile** is this post, if at all?

12:29

Please evaluate the following **randomly-selected** post.

These dumbass republicans have done nothing to fix our economy. Completely useless

3:32 PM · July 14, 2022 ·

How **engaging** is this post, if at all?

How **hostile** is this post, if at all?

Figure 3. Example of what respondents saw in the evaluation portion of the survey experiment. The image on the left is what control respondents would have seen. The images in the middle and far right were shown to treated respondents (i.e., those who were given a choice over the types of content they consumed). The middle image was shown to those who treated respondents who chose to view algorithmically-selected posts, and the image on the far right was show to those who chose to view randomly selected posts.

Our Sample

During our recruitment process (which took place from November 30th 2022 to February 11th 2023), we used CloudResearch's demographic filtering tools to specifically target U.S.-based MTurkers. We aimed to recruit 950 Democrats and 950 Republicans. If an MTurker fits our targeted demographic, then CloudResearch will post a "HIT request" (Mturk's version of a call for respondents) on the Mturker's feed. The Mturker can then choose whether to take our HIT based on whether they are interested in the task. To see a copy of the HIT description, please see Supplementary Information 6. Respondents were offered $1.82 for 10 minutes of their time.

After excluding respondents who did not complete the survey-experiment, those who used the same survey-completion code more than once, and those who were based out of the United States, we were left with 1669 respondents.

We then created quality flags to indicate whether a respondent sped through the survey (i.e., completed the survey in less than half the median time taken to complete the survey), whether their IP address was used by another respondent, whether they gave "junk" or non-sequitir responses to an open-ended survey question.

We eliminated respondents who failed two or more quality checks, leaving us with 1602 respondents. Among them, 49% were in the control condition, and 51% were in the treatment condition. Please see Supplementary Information 1 for a detailed breakdown of quality check failures. For a full demographic breakdown by condition, refer to Table 1 below.

Table 1. Comparative demographics of control and treated respondents. For standard balance table, please see SI 2.

	Control % (N = 781)	**Treated %** (N = 821)
Party		
Democrats	46.86	47.26
Republicans	44.17	45.19
Independents	7.94	7.06
Political Ideology		
Liberal	46.86	44.09
Conservative	44.17	46.89
Gender		
Women	59.41	59.68
Men	40.08	39.22
Other gender	0.51	0.73
Prefer not to disclose	NA	0.37
Race		
White	78.36	83.07
Not White	21.64	16.93
Education		
Bachelor's degree	53.91	58.10
No Bachelor's degree	46.09	41.90
Age		
18 - 24	4.48	4.87
25 - 34	25.48	27.65
35 - 44	30.86	28.26
45 - 54	18.95	19.85
55 - 64	12.55	12.06
65+	7.68	7.31

Results

Overall Effects of Choice on Platform Evaluation and Content Perception

In our first set of analyses we evaluate the overall effect of being offered a choice on respondents' evaluation of the platform and evaluation of the hostility of the posts that they see. Figure 4A reports overall satisfaction with the platform, significant at the $p = 0.1$ level, which appears to be driven by respondents' reported likelihood of recommending the platform to others ($p = 0.007$). While we see these positive evaluations of our platform (which would indicate there are benefits to offering choice to users), we do not see significant differences between those who were offered the choice to avoid toxic political content and those who were not on measures of average hostility of the posts they see (see Figure 4B).

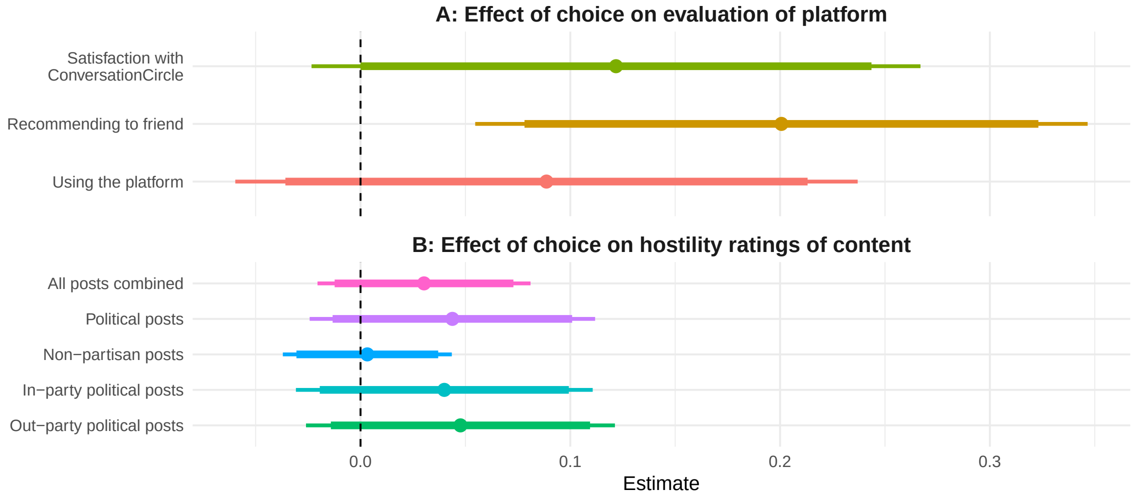


Figure 4. Full sample estimates of the effects of offering respondents the choice to avoid toxic political content. Panel A, above, reports overall satisfaction with the platform, significant at the $p = 0.1$ level, which appears to be driven by respondents' reported likelihood of recommending the platform to others ($p = 0.007$). While we see these positive evaluations of our platform (which would indicate there are benefits to offering choice to users), we do not see significant differences between those who were offered the choice to avoid toxic content and those who were on measures of average hostility of the posts they see (see Panel B). All the bars describe 95% and 90% confidence intervals. For full regression tables, please refer to Table SI 6 and Table SI 7 in the online supplement, respectively.

Who Chooses to Avoid Toxic Political Content?

One aspect of choice that this set of analyses does not capture is the effect of offering choice on users who *actually* would make the choice to avoid toxic content. As policy makers and regulators continue to investigate ways of improving social media's impact on its users, it becomes crucial to consider the actual adoption and uptake of proposed interventions. In other words, it is of particular significance to examine the effects of a new intervention on users who are truly exposed to it. Our study design allows us to evaluate the effect of being offered choice on those who would accept the choice to avoid toxic political content by comparing individuals who accepted the choice in the treatment arm with those who *would have* accepted the choice had they been offered it in the control arm.

To calculate this quantity of interest, we need to determine which individuals would have accepted the offer to see algorithmically-selected posts to avoid toxic political content, if they had been given a choice. Of the 821 respondents who were given the choice to avoid toxic political content, roughly 62% selected this option. Broadly, this suggests that most users will choose to avoid toxic political content if given the option to do so. Figure 5 reports the correlates of accepting the offer to avoid toxic political content, if given the choice. Looking at which respondents made that choice, we find that accepting the offer was strongly predicted by existing social media attitudes. Those satisfied with social media were more likely to accept the offer to see algorithmically-selected posts to avoid toxic political content, while those who believe social media has an online bullying problem were less likely to choose randomly-selected content. In contrast, political and demographic attitudes are not predictive of making the choice.

To understand the treatment effect for those likely to choose to avoid toxic political content, we performed a logistic regression on the treatment group to explore which factors influenced their choice. The results, depicted as odds ratios in Figure 5, highlight the relationship between various predictors and the likelihood of choosing to filter out toxic political content. We then used this model to estimate the propensity of individuals in the control group to make a similar choice, had they been given the opportunity. To make this comparison we calculate the difference in average outcomes between individuals in the treated arm who accepted this choice and the average weighted outcomes of individuals in the control arm (where the weights are the predicted probabilities of choosing to view less toxic political content). This comparison identifies the treatment effect among those who would choose to avoid toxic political content if given the option under the assumption that being given the choice does not influence the option chosen.

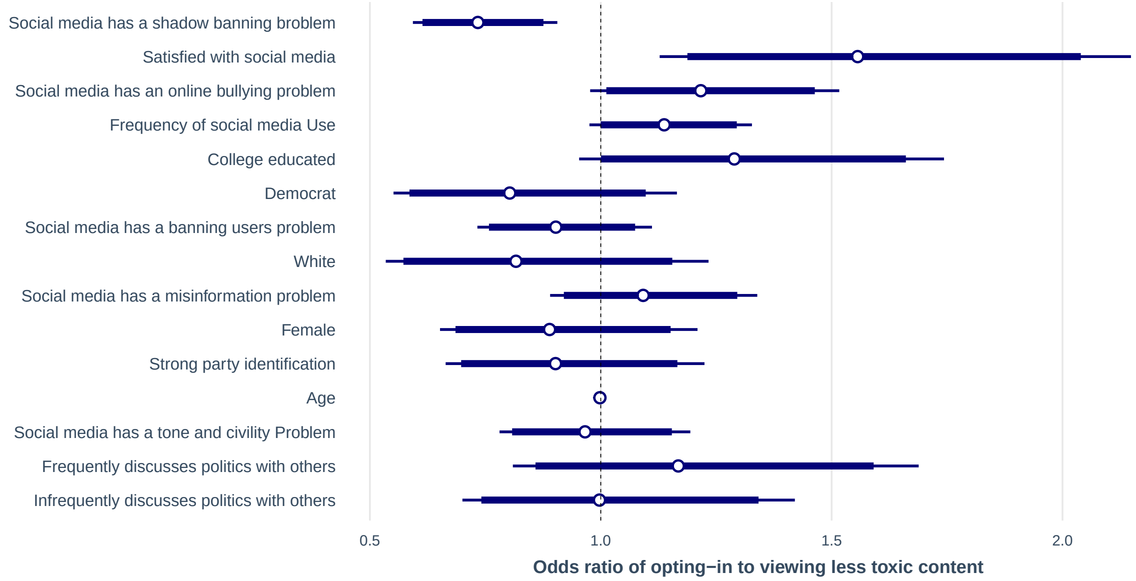


Figure 5. Odds ratios of accepting offer to see less toxic political content. Those who believe that social media has a shadow-banning problem were far less likely to choose to see less toxic content when given the option (OR $= 0.734$, $p = 0.004$). Those who are more satisfied with their experience on social media were far more likely to accept (OR $= 1.556$, $p = 0.007$). This may pose a challenge for social media platforms aiming to enhance user experiences for dissatisfied users, as they may be less receptive to changes or new features. Believing that there is an issue with online bullying and using social media more frequently had a positive relationship with choosing to see less toxic content, although these variables approached but did not reach statistical significance (OR $= 1.217$, $p = 0.080$ and OR $= 1.137$, $p = 0.099$, respectively). Odds ratios are computed from a multivariate logistic regression with choice as the binary outcome regressed on all of the listed predictors. For the full regression table, please refer to Supplementary Table SI 8 in the online supplement.

Effects of Choice on Those Who Opt for Content Filtering

Figure 6 shows the effect of being offered the choice to avoid toxic political content among those who would make that choice. When comparing the effects of choice among respondents in treatment who accepted the offer to see less toxic political content to respondents in control who would have made the same choice (Figure 6B), we find that those who were actually given a choice (i.e., were in the treatment condition) were more likely to rate all social media posts as hostile ($p = 0.007$) compared to those respondents in control who would have made the same choice, even though all respondents, regardless of condition, saw the same content. They were more like to rate political and non-partisan posts as hostile ($p = 0.015$ and $p = 0.060$, respectively). Furthermore, when we break the political posts into in- and out-party, we find that, again, respondents who were offered the choice and accepted the option to see less toxic political content were more likely to rate such posts as hostile ($p = 0.033$ and $p = 0.014$, respectively). These results suggest that effect of offering choice can lead to negative attitudes about platform content. Paradoxically, however, the same users who rated the content negatively expressed greater satisfaction with ConversationCircle ($p = 0.045$). This effect appears to be driven by respondent's propensity to recommend the platform to others ($p = 0.001$) (see Figure 6A).

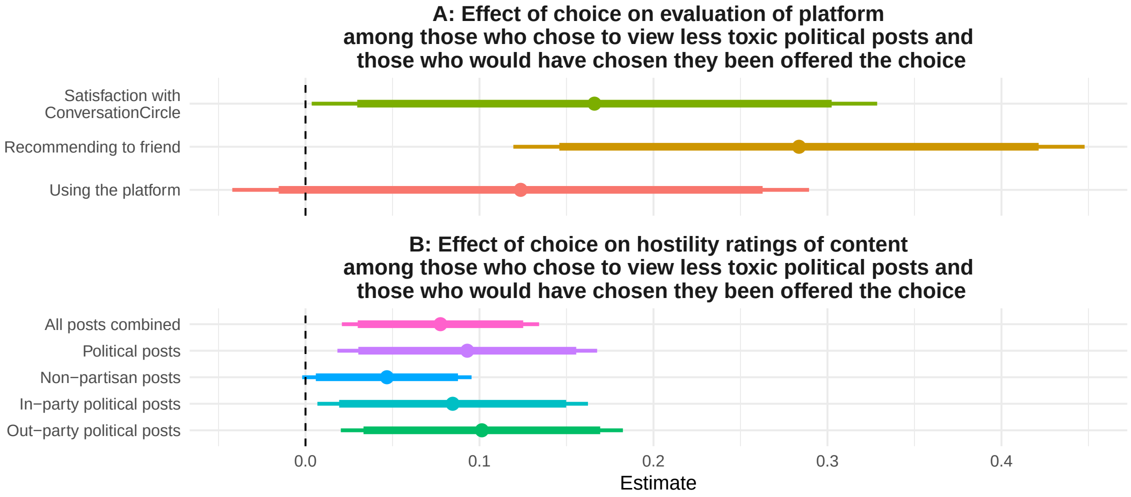


Figure 6. Effects of choice among respondents who chose to view less toxic political content and those in the control condition who would have chosen similarly, had they been offered the choice. Respondents given the choice rated all posts as more hostile ($p = 0.007$), with this trend persisting for political and non-partisan content ($p = 0.015$ and $p = 0.060$, respectively), as well as in- and out-party posts ($p = 0.033$ and $p = 0.014$, respectively), even though all respondents, regardless of condition saw the same posts. Interestingly, although these users perceived the platform's posts as hostile, these respondents expressed greater satisfaction with the platform ($p = 0.045$), primarily due to their increased willingness to recommend it ($p = 0.001$). All the bars describe 95% and 90% confidence intervals. All the bars describe 95% and 90% confidence intervals. For full regression tables, please refer to Table SI 9 and Table SI 10 in the online supplement, respectively.

Our findings reveal that giving users choice over content exposure may create unintended consequences. Specifically, respondents who chose algorithmically-filtered content rated posts as more hostile than those in the control group, despite viewing identical content. This suggests that offering users the choice to avoid toxic content may heighten their expectations— when we told respondents the algorithm would prioritize positive content, they appeared more sensitive to any content that failed to meet this standard.

Follow-up Study: Testing Effects with No Negative Content

To test whether these negative evaluations stemmed from exposure to toxic political content, we conducted a follow-up study using the same design but excluding all negative political posts. This modification offered the strongest possible test of whether choice alone could improve users' content evaluations.

The follow-up study revealed two key findings. First, being offered choice once again increased satisfaction with the platform (Figure 7). Respondents who chose to view less toxic political posts had an estimated treatment effect of 0.270 ($p = 0.004$) for overall satisfaction with ConversationCircle, 0.298 ($p = 0.003$) for recommending that platform to others, and 0.300 ($p = 0.003$) for using the platform themselves.

Even eliminating any negative content, we still find that users who were offered choice did not rate the posts more favorably than those not offered choice. The estimated treatment effect on the hostility ratings of the platform's posts was 0.082 ($p = 0.054$). A similar pattern was found looking at the specific types of content, including political posts, non-partisan posts, in-party, and out-party political posts (hostility effects of 0.110, $p = 0.059$; 0.054, $p = 0.185$; 0.101, $p = 0.096$; and 0.120, $p = 0.081$, respectively). These results suggest that the mere absence of negative content does not automatically improve perceptions of the content. A full overview of the results of our follow-up study can be found in Supplementary Information 2.

Comparing Effects Across Studies: The Role of Content Context

To more explicitly test the difference in treatment effects across the two studies, we estimated a model that combined the data from the two studies and included an interaction term between the treatment indicator and the study ID. Importantly, this requires an assumption that both studies are generalizable to the same population, which is plausible in our setting as both studies were run on the MTurk platform, where the second study excluded those who participated in the first study. Since respondents in both studies were exposed to four of the same positive political posts, we used the average hostility score of those posts as the outcome (please refer to Supplementary table SI 16 in the online supplement for the full regression table). We see a statistically significant difference between studies (posts in the second study, where all negative political posts were excluded, are rated as more hostile), a statistically significant treatment effect across both studies, and the interaction term was effectively zero and not statistically significant.

In other words, when no negative posts were present (i.e., in our follow-up study), respondents rated the positive posts as more hostile than respondents in our original study (who had seen a mix of both positive and negative posts). This suggests a potential "contrast" effect: when users see a mix of positive and negative content, they might rate the positive posts as less hostile because they have a point of comparison—the negative content. This implies that the evaluation of content's hostility is not static. Instead, users may judge content relative to the context in which it is viewed. Once again, these findings indicate offering choice might improve overall evaluations of the platform, but it might also change expectations about the particular content that could be encountered on the platform.

Discussion

Our study builds on a growing but still limited body of work on user autonomy in algorithmically-driven spaces, addressing questions raised by both policymakers and social scientists about the implications of user-based content moderation (Rathje et al., 2024; Jhaver et al., 2023; Jhaver and Zhang, 2023; Gorwa et al., 2020; Haimson et al., 2021). Our findings show how the mere promise of algorithmic control fundamentally shapes how users interpret social media content. When offered the choice to filter out toxic political content, respondents who opted for this choice rated all posts as more hostile than equivalent respondents not offered any choice—despite seeing identical content. This effect did not dissipate in our follow-up study where we removed all negative content, suggesting that the mere promise of algorithmic control shapes how users interpret content, regardless of the content's actual nature.

This finding challenges straightforward assumptions about the benefits of user autonomy on social media plat-

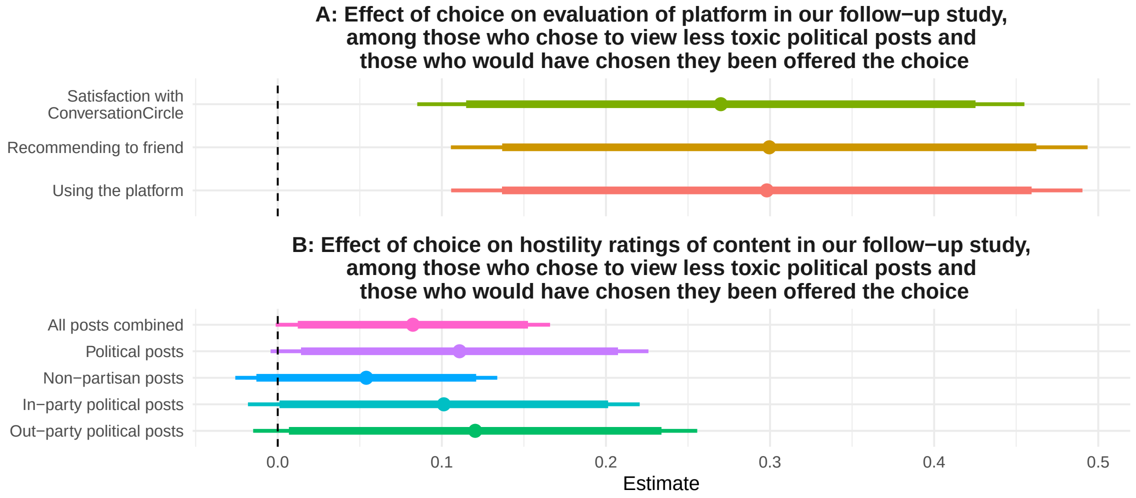


Figure 7. This figure shows the effects of providing choice to respondents who chose to view less toxic political content and controls who would have also made that choice had they been offered the choice in our follow-up study. Panel A shows that respondents who were offered choice were more likely to be satisfied with the platform, overall ($p = 0.004$). This effect is driven by the effect of choice on respondents' reported likelihood of using and recommending the platform to others ($p = 0.002$ and $p = 0.003$, respectively). In Panel B, we find that offering respondents choice over the type of content they consumed did not significantly impact hostility ratings at the 0.05 level. Overall, the point estimate of the treatment effect for those who were given a choice to view less toxic political content relative to respondents in control who *would have* chosen to see less toxic political content had they been offered the choice on hostility ratings was 0.082 ($p = 0.054$). When we look at the mean hostility ratings by type of content, we find that the respondents who chose to view less toxic posts had treatment effect point estimates on hostility ratings of 0.110 on political posts ($p = 0.059$), 0.054 on non-partisan posts ($p = 0.185$), as well as 0.101 on in-party ($p = 0.096$) and 0.120 on out-party ($p = 0.081$) political posts. All the bars describe 95% and 90% confidence intervals. For full regression tables, please refer to Table SI 14 and Table SI 15 of the online supplement, respectively.

forms. Rather than simply improving user experiences, the promise of control seems to heighten user expectations, making them more critical of content that falls short of these elevated standards. This dynamic reflects Seaver's (2019) broader observation about the fundamental challenge of algorithmic systems: it's not just about knowing how to technically implement a specific feature, but about the complex cultural work required to translate socio-cultural concepts into computational code. Just as Seaver describes how features like "affinity scores" require careful interpretation and formalization, our findings show how platforms must grapple not only with defining and operationalizing "toxicity," but with the reality that users themselves bring their own interpretations of what constitutes toxic content. Our findings suggest that merely promising to filter toxic political content potentially activates users' own standards and expectations, leading them to be more critical of content that might not align with their personal definitions of what should be filtered. Yet paradoxically, these same users who perceived greater hostility also reported higher satisfaction with the platform itself. This tension echoes Eslami et al.'s 2015 finding that while users initially react negatively to discovering the existence of algorithmic filtering, having agency in the process ultimately increases satisfaction. As the authors argue, making it possible for users to actively engage with algorithmic systems gives them "an important sense that they are not controlled by an algorithm but are a part of one" (p. 9). Our results similarly suggest that the offering of control creates value for our respondents even when it leads to more negative interpretations of the content they viewed.

Our comparison in hostility ratings between the two experiments revealed another layer of complexity in how users evaluate content. When we removed all negative posts in our follow-up study, respondents rated the positive posts as more hostile than respondents in our first study who saw a mix of positive and negative content. This "contrast effect" suggests that content evaluation is not absolute but relative—respondents may need points of comparison to benchmark what constitutes hostile content.

These findings have important implications for platform governance and strategies for content moderation. While platforms increasingly offer users control over their feeds—from TikTok's "For You" preferences to Meta's content filtering options to Bluesky Social's fully customizable algorithmic feeds—our results suggest these features may inadvertently raise user expectations in ways that make them more critical of all content. This creates a challenging paradox for platforms: while offering control may increase overall user satisfaction, it might also lead to more negative perceptions of the very content these controls are meant to improve.

This dynamic strongly aligns with the notion of algorithmic imaginaries, which describes how users develop their own theories about how algorithms work and adjust their perceptions accordingly (Bucher, 2017). Just as Bucher found that Facebook users developed specific theories about how to "game" the algorithm to maintain visibility, our respondents may have developed specific strategies for engaging with our platform. When promised less toxic political content, respondents who rated content as more hostile may have been attempting to "train" the algorithm to better recognize what they deem to be “toxic,” similar to how Eslami et al. (2016) found that users deliberately interact with content to shape future algorithmic recommendations. This suggests that content curation is co-produced by users and algorithms– users don’t just passively receive filtered content but actively try to influence how algorithms interpret and classify content.

Moreover, our findings reveal a potential challenge for platform reform efforts: satisfied social media users in our sample were more likely to choose to try out the new content filtering option, while dissatisfied users avoided it. This self-selection creates a paradox where new control features may fail to reach the very users most frustrated with social media toxicity. Platforms hoping to improve user experiences through increased autonomy may thus find their efforts undermined when their most critical users reject these new features.

Our findings open new directions for research on user control and algorithmic systems. Beyond autonomy, future studies should explore how different forms of control—such as ranking content, modifying algorithmic priorities, or providing direct feedback—affect user experiences and engagement across platforms. Comparing the highly tailored algorithms offered by platforms like Bluesky with the more traditional systems of X (formerly Twitter) could reveal how varying levels of control influence user satisfaction, trust, and perceptions of content.

Another promising avenue involves examining how users interpret and respond to algorithmic systems differently across platforms, particularly how these interactions reinforce or challenge algorithmic imaginaries. Social media users often develop creative strategies to navigate perceived algorithmic suppression. For example, TikTok users, believing that content mentioning words like “death” or “suicide” might be suppressed, have adopted alternative terms like “unalived,” a workaround that has since entered broader cultural slang. During a viral discussion about drones over New Jersey, users substituted the term “Dior bags” to avoid perceived penalties, showing how users actively reshape their communication strategies en masse. In some cases, these strategies extend to altruistic goals: users fundraising for a refugee family might create videos focused on innocuous topics, such as a steel water bottle, and encourage viewers to engage with the video by commenting or searching for related terms, believing this activity will amplify its reach. Future research could investigate how such imaginaries evolve over time and whether greater algorithmic transparency - like those offered by Bluesky - can lead to more accurate understandings of platform mechanics or if these imaginaries persist despite the transparency.

Lastly, understanding how algorithms and content moderation influence offline behaviors and social norms could shed light on the broader societal impact of platform design. Movements like #BlackLivesMatter and #MeToo show how the algorithmic amplification (which led to the virality) of hashtags can spark offline protests and drive policy changes. Similarly, the widespread adoption of alternative language, such as “unalived,” and the creative use of algorithmic strategies for humanitarian causes reveal how digital practices shape cultural norms and behaviors beyond the online space. By exploring the intersections between algorithmic systems and real-world actions, future research can provide deeper insights into how user control and platform design influence both individual behavior and collective outcomes.

While our study provides valuable insights into how offering users more control affects their experiences and interpretation of content, there are several important limitations. First, our experimental design captured only static, one-time interactions with content rather than the dynamic nature of real social media use. Respondents were told they were evaluating a new platform ("ConversationCircle") and so might have approached content more analytically than they would on established social media sites. Similarly, respondents’ awareness of taking part in a study may have prompted more deliberate content evaluation than occurs in natural settings. Second, our sample characteristics may also limit generalizability. MTurk workers may have greater technical literacy and algorithm awareness than average social media users.

Third, our twelve fabricated posts cannot capture the full range of content users encounter on social media. By focusing specifically on toxic political content, we may have missed how users interpret and react to other types of toxic content online. Moreover, we presented all posts anonymously, blurring names and profile pictures. Users might evaluate content differently when they can see who posted it– whether from friends, public figures, or known organizations. This anonymity may have stripped away important social context that typically influences how users interpret and react to social media content.

Finally, our reliance on self-reported hostility ratings may not capture unconscious reactions to content. We also measured platform satisfaction by asking respondents whether they would hypothetically use or recommend ConversationCircle, rather than observing actual user behavior. These hypothetical measures may not accurately reflect how users would engage with the platform in practice. These constraints point to opportunities for future research using more naturalistic settings, diverse content types, and behavioral measures of user response.

Our findings make important contributions to ongoing debates about democracy and the public sphere in algorithmically-mediated spaces by demonstrating how platform mechanisms fundamentally shape not just what content users see, but how they interpret and evaluate that content. This has significant implications for democratic discourse online, as users' interpretations of content toxicity and hostility may be influenced as much by platform promises and controls as by the content itself.

Author contributions statement

F.A., S.H., C.B., and A.V. conceived of the experiment(s), F.A. conducted the experiment(s), F.A. and G.T. analysed the results, F.A. created the figures and tables. All authors reviewed and contributed to the manuscript.

Competing Interests: Authors declare they have no competing interests.

Data and Code Availability

Anonymized replication code and data will be made publicly available by the authors at this link upon the publishing of the manuscript: [REDACTED].

Funding

Google Jigsaw Research Award in the form of unrestricted gifts.

References

2022. “Control Your Instagram Feed | Instagram Blog”. https://about.instagram.com/blog/tips-and-tricks/control-your-instagram-feed.

Ahmed, Aaliyah. 2024, November. “What is Bluesky Social and who owns the X competitor?”. *The Times*.

Allyn, Bobby. 2021, October. “Here are 4 key points from the Facebook whistleblower’s testimony on Capitol Hill”. *NPR*.

Auxier, Brooke. 2020. “64% of Americans say social media have a mostly negative effect on the way things are going in the U.S. today. pew research center”. Pew Research Center https://pewrsr.ch/3dsV7uR.

Bailey, Michael A, Daniel J Hopkins, and Todd Rogers. 2016. “Unresponsive and unpersuaded: The unintended consequences of a voter persuasion effort”. *Political Behavior* 38 : 713–746.

Beknazar-Yuzbashev, George, Rafael Jiménez Durán, Jesse McCrosky, and Mateusz Stalinski. 2022. “Toxic Content and User Engagement on Social Media: Evidence from a Field Experiment”. *SSRN Electronic Journal*.

Bluesky. 2023. “Algorithmic Choice with Custom Feeds”.

Bode, Leticia. 2016. “Pruning the news feed: Unfriending and unfollowing political content on social media”. *Research & Politics* 3 .

Brady, William J, Killian McLoughlin, Tuan N Doan, and Molly J Crockett. 2021. “How social learning amplifies moral outrage expression in online social networks”. *Science Advances* 7 (33).

Bucher, Taina. 2017, January. “The algorithmic imaginary: Exploring the ordinary affects of Facebook algorithms”. *Information, Communication & Society* 20 (1): 30–44.

C-Span. 2021. “Senate Hearing on Social Media Algorithms | C-SPAN.org”.

Cheng, Justin, Michael Bernstein, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2017. “Anyone Can Become a Troll: Causes of Trolling Behavior in Online Discussions”. In *Proceedings of the 2017 ACM Conference on Computer Supported Cooperative Work and Social Computing*, pp. 1217–1230. ACM.

Cole, Amato. 2024, November. “Bluesky vs. Threads: A Comparison of New Social Media Players”. *Yahoo! Tech*.

Dember, William N., Traci L. Galinsky, and Joel S. Warm. 1992. “The role of choice in vigilance performance”. *Bulletin of the Psychonomic Society* 30 (3): 201–204.

Dooms, Simon, Toon De Pessemier, and L. Martens. 2014. “Improving IMDb Movie Recommendations with Interactive Settings and Filters”. In *RecSys Posters*.

Eslami, Motahhare, Karrie Karahalios, Christian Sandvig, Kristen Vaccaro, Aimee Rickman, Kevin Hamilton, and Alex Kirlik. 2016, May. “First I "like" it, then I hide it: Folk Theories of Social Feeds”. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, San Jose California USA, pp. 2371–2382. ACM.

Eslami, Motahhare, Aimee Rickman, Kristen Vaccaro, Amirhossein Aleyasen, Andy Vuong, Karrie Karahalios, Kevin Hamilton, and Christian Sandvig. 2015, April. “"I always assumed that I wasn’t really that close to [her]": Reasoning about Invisible Algorithms in News Feeds”. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, Seoul Republic of Korea, pp. 153–162. ACM.

Gerrard, Ysabel. 2018, December. “Beyond the hashtag: Circumventing content moderation on social media”. *New Media & Society* 20 (12): 4492–4511.

Gillespie, Tarleton. 2010, May. “The politics of ‘platforms”’. *New Media & Society* 12 (3): 347–364.

Gorwa, Robert, Reuben Binns, and Christian Katzenbach. 2020, January. “Algorithmic content moderation: Technical and political challenges in the automation of platform governance”. *Big Data & Society* 7 (1): 2053951719897945.

Haimson, Oliver L., Daniel Delmonaco, Peipei Nie, and Andrea Wegner. 2021, October. “Disproportionate Removals and Differing Content Moderation Experiences for Conservative, Transgender, and Black Social Media Users: Marginalization and Moderation Gray Areas". *Proc. ACM Hum.-Comput. Interact.* 5 (CSCW2): 466:1–466:35.

Harris, Elizabeth, Steve Rathje, Claire E. Robertson, and Jay J. Van Bavel. 2023, January. "The SPIR Framework of Social Media and Polarization: Exploring the Role of Selection, Platform Design, Incentives, and Real-World Context". *International Journal of Communication (19328036)* 17 : 5316–5335.

Jhaver, Shagun and Amy Zhang. 2023. "Do Users Want Platform Moderation or Individual Control? Examining the Role of Third-Person Effects and Free Speech Support in Shaping Moderation Preferences".

Jhaver, Shagun, Alice Qian Zhang, Quan Ze Chan, Nikhila Natarajan, Ruotong Wang, and Amy X. Zhang. 2023. "Personalizing content moderation on social media: User perspectives on moderation choices, interface design, and labor". *Proc. ACM Hum.-Comput. Interact.*.

Knijnenburg, Bart P., Svetlin Bostandjiev, John O'Donovan, and Alfred Kobsa. 2012. "Inspectability and control in social recommenders". In *Proceedings of the Sixth ACM Conference on Recommender Systems - RecSys '12*, pp. 43. ACM Press.

Langer, Ellen J. 1975. "The illusion of control". *Journal of Personality and Social Psychology* 32 : 311–328.

Langer, E. J. and J. Rodin. 1976. "The effects of choice and enhanced personal responsibility for the aged: A field experiment in an institutional setting". *Journal of Personality and Social Psychology* 34 (2): 191–198.

McNee, Sean M., Shyong K. Lam, Joseph A. Konstan, and John Riedl. 2003. "Interfaces for eliciting new user preferences in recommender systems". In P. Brusilovsky, A. Corbett, and F. de Rosis (Eds.), *User Modeling 2003*, Lecture Notes in Computer Science, pp. 178–187. Springer.

Meta. 2020, January. "Expanded Transparency and More Controls for Political Ads".

Nanz, Andreas and Jörg Matthes. 2022. "Democratic consequences of incidental exposure to political information: A meta-analysis". *Journal of Communication* 72 (3): 345–373.

Papacharissi, Zizi. 2008. "The virtual sphere 2.0: The internet, the public sphere, and beyond". In *Routledge Handbook of Internet Politics*. Routledge.

Persily, Nathaniel and Joshua A. Tucker (Eds.)2020. *Social Media and Democracy: The State of the Field, Prospects for Reform*. SSRC Anxieties of Democracy. Cambridge University Press.

Rathje, Steve, Claire Robertson, William J Brady, and Jay J Van Bavel. 2024. "People Think That Social Media Platforms Do (but Should Not) Amplify Divisive Content". *Perspectives on Psychological Science*.

Rathje, Steve, Jay J. Van Bavel, and Sander van der Linden. 2021. "Out-group animosity drives engagement on social media". *Proceedings of the National Academy of Sciences* 118 (26).

Reynolds, Glenn H. 2007. *An Army of Davids: How Markets and Technology Empower Ordinary People to Beat Big Media, Big Government, and Other Goliaths*. Nelson Current.

Sato, Mia. 2024, August. "TikTok is adding new ways to fine-tune your For You Page algorithm - The Verge". *The Verge*.

Seaver, Nick. 2019, May. "Knowing Algorithms". In *digitalSTS: A Field Guide for Science & Technology Studies*. Princeton University Press.

Van Dijck, José, Thomas Poell, and Martijn De Waal. 2018, October. *The Platform Society*, Volume 1. Oxford University Press.

Xiao, Bo and Izak Benbasat. 2007. "E-commerce product recommendation agents: Use, characteristics, and impact". *MIS Quarterly* 31 (1): 137–209.

Supplementary Information

“Supplemental Appendix for: Experiments offering social media users the choice to avoid toxic political content"

Fatima Alqabandi[1,2], **Graham Tierney**[2], **Christopher Bail**[*1,3,4], **Sunshine Hillygus**[3,4], **and Alexander Volfovsky**[2,5]

[1]Duke University, Department of Sociology

[2]Duke University, Department of Statistical Science

[3]Duke University, Department of Political Science

[4]Duke University, Sanford School of Public Policy

[5]Duke University, Department of Computer Science

[*]christopher.bail@duke.edu

Supplementary Information 1 Quality checks and standard balance table

SI 1.1 Quality checks

Table SI 1 presents the breakdown of participants by the number of quality checks they failed.

Table SI 1. Comparison of Quality Check Failure Rates Between Control and Treatment Groups

Number of quality checks failed	Control	Treatment
0	593	695
1	188	126
2	45	14
3	6	2

We eliminated those who failed 2 or more quality checks. This left us with 1602 participants, 49% of whom were in the control condition, and 51% in the treatment condition.

Table SI 2. Balance between participants in control and treatment conditions

Variable	Control ($N = 781$)	Treatment ($N = 821$)	*P*-value
Age: Over 35 y.o.	0.66	0.65	0.52
Race: White	0.78	0.83	0.02
Educ: Bachelors or more	0.54	0.58	0.09
Gender: Female	0.59	0.60	0.91
PID: Democrat	0.47	0.47	0.87
PID: Republican	0.44	0.45	0.68

Note: Proportions shown in “Control" and “Treatment" columns. P-values from two-tailed tests. Independents are excluded.

Supplementary Information 2 Follow-up study

For our follow-up study, we relaunched our original experiment with one major difference: there were no negative content for participants to evaluate. For a breakdown of our sample characteristics, please refer to Supplementary Table SI 3 below. We used the exact same methods and analysis plan for this study as we did in our original (as described in the Methods section of the corresponding paper).

Much like we did in the Results section of the original study, we first begin by presenting the overall effects of giving participants choice, followed by the effects of offering choice among participants who tend to opt-in to viewing less toxic content.

Table SI 3. Follow-up Study: Demographics of our control sample on the left, and our treated sample on the right

Treatment	N	%	**Control**	N	%
Party			**Party**		
Democrats	260	48.96	Democrats	257	46.14
Republicans	222	41.81	Republicans	249	44.70
Independents	40	7.53	Independents	42	7.54
Political ideology			**Political ideology**		
Liberal	253	47.65	Liberal	239	42.91
Conservative	225	42.37	Conservative	258	46.32
Gender			**Gender**		
Women	330	62.15	Women	328	58.89
Men	197	37.10	Men	225	40.39
Other gender	2	0.38	Other gender	3	0.54
Prefer not to disclose	NA	NA	Prefer not to disclose gender	1	0.18
Race			**Race**		
White	446	83.99	White	464	83.30
Not White	85	16.01	Not White	92	16.52
Education			**Education**		
Bachelor’s degree	292	54.99	Bachelor’s degree	323	57.99
No Bachelor’s degree	238	44.82	No Bachelor’s degree	234	42.01
Age			**Age**		
18 - 24	37	6.97	18 - 24	42	7.54
25 - 34	180	33.90	25 - 34	147	26.39
35 - 44	148	27.87	35 - 44	177	31.78
45 - 54	85	16.01	45 - 54	98	17.59
55 - 64	54	10.17	55 - 64	67	12.03
65+	27	5.08	65+	26	4.67
Total	531	100.00	**Total**	557	100.00

SI 2.1 Effects of choice

We first looked at the overall effects of giving participants choice (Supplementary Fig. SI 1). We found that there is a strong positive effect of giving users choice over the content they consume on participants' overall satisfaction with our simulated platform ($p = 0.008$). This effect was driven by participants reporting higher likelihood of using and recommending the platform ($p = 0.006$ and $p = 0.012$, respectively) (Supplementary Fig. SI 1A). We find no significant differences in hostility ratings of posts between participants who were given a choice to avoid toxic political content and those who were not (Supplementary Fig. SI 1B).

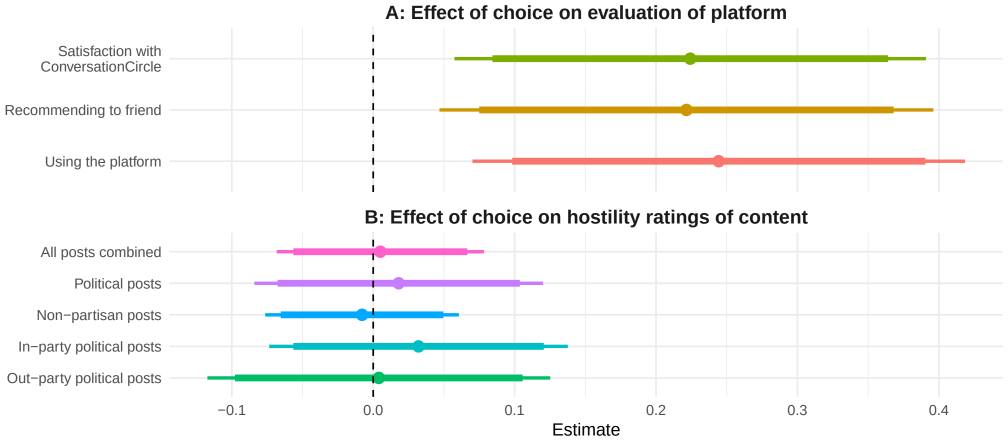


Figure SI 1. Effects of offering participants the choice to avoid toxic political content our follow-up study (where all negative content had been removed). We find that there is a strong positive effect of giving users choice over the content they consume on participants' overall satisfaction with our simulated platform ($p = 0.008$), and their reported likelihood of using and recommending the platform ($p = 0.006$ and $p = 0.012$, respectively). We find no significant differences in hostility ratings of posts between participants who were given a choice to avoid toxic political content and those who were not. All the bars describe 95% and 90% confidence intervals.

SI 2.2 Effects of choice on those who opt-in

Similar to the proportions in our original experiment, 64% of treated participants opted in to viewing less toxic posts, with 36% opting-out.

Looking at the effects of treatment on those who would have opted-in to viewing less toxic, algorithmically-selected content had they been offered a choice (Supplementary Fig. SI 2), we find that those who were given a choice were far more likely to be satisfied with our simulated platform ($p = 0.004$), and to use and recommend ConversationCircle to others ($p = 0.002$ and $p = 0.003$, respectively).

Overall, the point estimate of the treatment effect for those who were given a choice to view less toxic political content relative to participants in control who *would have* opted-in had they been offered the choice on hostility ratings was 0.082 ($p = 0.054$). When we look at the mean hostility ratings by type of content, we find that the participants who opted-in to view less toxic posts had treatment effect point estimates on hostility ratings of 0.110 on political posts ($p = 0.059$), 0.054 on non-partisan posts ($p = 0.185$), as well as 0.101 on in-party ($p = 0.096$) and 0.120 on out-party ($p = 0.081$) political posts.

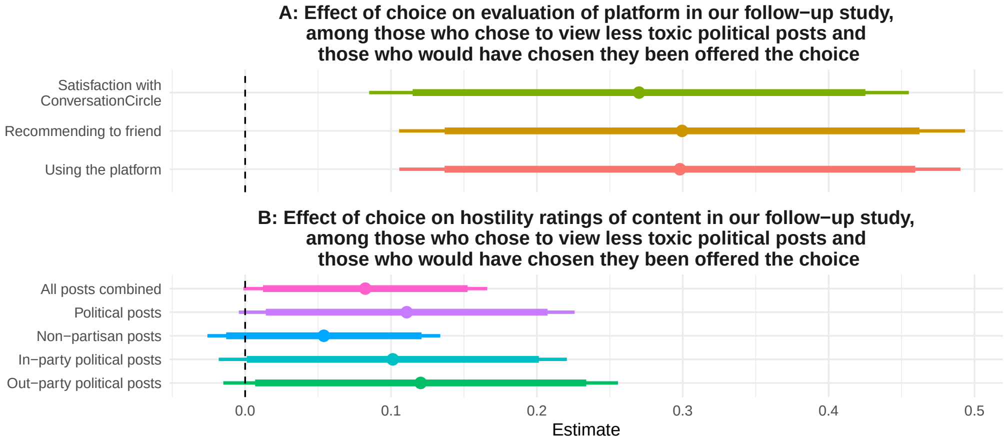


Figure SI 2. Effects of choice among participants who opted-in to viewing less toxic political content and controls who would have opted-in had they been offered the choice in our follow-up study. Panel A shows that participants who were offered choice were more likely to be satisfied with the platform, overall ($p = 0.004$). This effect is driven by the effect of choice on participants' reported likelihood of using and recommending the platform to others ($p = 0.002$ and $p = 0.003$, respectively). In Panel B, we find that offering participants choice over the type of content they consumed did not significantly impact hostility ratings at the 5% level. Overall, the point estimate of the treatment effect for those who were given a choice to view less toxic political content relative to participants in control who *would have* opted-in had they been offered the choice on hostility ratings was 0.082 ($p = 0.054$). When we look at the mean hostility ratings by type of content, we find that the participants who opted-in to view less toxic posts had treatment effect point estimates on hostility ratings of 0.110 on political posts ($p = 0.059$), 0.054 on non-partisan posts ($p = 0.185$), as well as 0.101 on in-party ($p = 0.096$) and 0.120 on out-party ($p = 0.081$) political posts. All the bars describe 95% and 90% confidence intervals.

Supplementary Information 3 Treatment effects among opt-outs

Below, we present the effects of offering choice on participants who would have opted-out in our original study and the follow-up.

SI 3.1 Treatment effects among opt-outs in original study

We find that offering choice to the types of people who would opt-out had a positive but statistically non-significant effect on users' satisfaction with ConversationCircle, their likelihood of recommending it to a friend, and their usage of the platform.

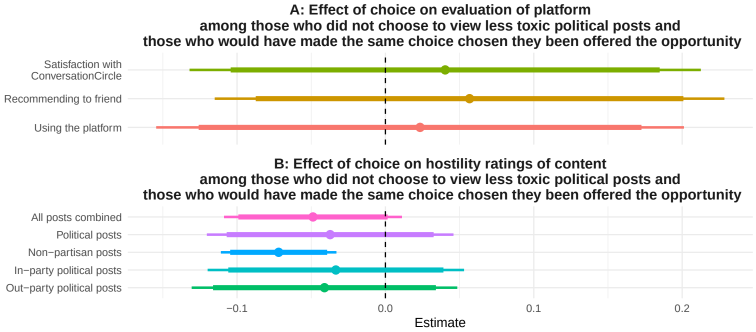


Figure SI 3. Panel A: Offering choice to people who would opt-out had a positive but statistically non-significant effect on users' satisfaction with ConversationCircle, their likelihood of recommending it to a friend, and their usage of the platform. Panel B: We find that offering said users the choice to avoid toxic political content results in them rating non-partisan posts as less hostile than users who would have opted-out but were not given the option to do so (p < .01). We saw no significant treatment effects for hostility ratings of the posts, overall, nor of political posts, in-party political posts, or out-party political posts. Full regression tables are reported below.

Table SI 4. Regression table for Figure SI 3A

	Dependent variable:		
	Satisfaction with ConversationCircle	Recommending to friend	Using the platform
Treatment condition	0.040	0.023	0.057
	(0.088)	(0.091)	(0.088)
Democrat	−0.335***	−0.257**	−0.411***
	(0.115)	(0.119)	(0.115)
College educated	−0.193**	−0.186**	−0.139
	(0.089)	(0.092)	(0.089)
Strong party ID	0.315***	0.263***	0.253***
	(0.091)	(0.094)	(0.091)
Age	0.009**	0.010***	0.009***
	(0.004)	(0.004)	(0.004)
Female	0.255***	0.177*	0.273***
	(0.092)	(0.095)	(0.092)
White	−0.128	−0.028	−0.135
	(0.122)	(0.126)	(0.122)
Frequency of SM use	0.092**	0.130***	0.056
	(0.040)	(0.042)	(0.040)
Satisfied with SM	0.304***	0.333***	0.336***
	(0.095)	(0.098)	(0.095)
Believe there is:			
misinformation problem on SM	−0.036	−0.046	−0.021
	(0.058)	(0.060)	(0.058)
a banning problem on SM	0.137**	0.129**	0.155**
	(0.063)	(0.065)	(0.062)
shadow-banning problem on SM	−0.077	−0.050	−0.071
	(0.064)	(0.066)	(0.064)
an onine bullying problem on SM	0.132**	0.145**	0.081
	(0.064)	(0.066)	(0.064)
a civility problem on SM	−0.115*	−0.173***	−0.076
	(0.061)	(0.063)	(0.061)
Frequently talk to others about politics	−0.053	−0.111	−0.071
	(0.109)	(0.113)	(0.109)
Infrequently talk to others about politics	−0.490***	−0.536***	−0.431***
	(0.105)	(0.109)	(0.105)
Constant	−1.245***	−1.687***	−1.515***
	(0.374)	(0.386)	(0.373)
Observations	1,084	1,084	1,084
R^2	0.107	0.104	0.105
Adjusted R^2	0.093	0.091	0.091
Residual Std. Error (df = 1067)	1.067	1.102	1.064
F Statistic (df = 16; 1067)	7.969***	7.768***	7.810***

Note: *p<0.1; **p<0.05; ***p<0.01

Table SI 5. Regression table for Figure SI 3B

	Dependent variable:				
	Mean hostility ratings				
	All posts combined	Political posts	Non-partisan posts	In-party political posts	Out-party political posts
Treatment condition	-0.049	-0.037	-0.072^{***}	-0.033	-0.041
	(0.031)	(0.042)	(0.020)	(0.044)	(0.046)
Democrat	0.095^{**}	0.129^{**}	0.026	0.090	0.168^{***}
	(0.040)	(0.056)	(0.026)	(0.058)	(0.060)
College educated	0.091^{***}	0.129^{***}	0.016	0.106^{**}	0.153^{***}
	(0.031)	(0.043)	(0.020)	(0.045)	(0.046)
Strong party ID	-0.040	-0.065	0.010	-0.037	-0.094^{**}
	(0.032)	(0.044)	(0.021)	(0.046)	(0.047)
Age	-0.002^{*}	-0.002	-0.003^{***}	-0.001	-0.002
	(0.001)	(0.002)	(0.001)	(0.002)	(0.002)
Female	-0.053^{*}	-0.056	-0.046^{**}	-0.038	-0.075
	(0.032)	(0.044)	(0.021)	(0.046)	(0.048)
White	-0.096^{**}	-0.116^{**}	-0.056^{**}	-0.109^{*}	-0.124^{*}
	(0.042)	(0.059)	(0.028)	(0.061)	(0.063)
Frequency of SM use	-0.012	-0.019	0.001	-0.023	-0.014
	(0.014)	(0.019)	(0.009)	(0.020)	(0.021)
Satisfied with SM	0.013	0.047	-0.055^{**}	0.056	0.039
	(0.033)	(0.046)	(0.021)	(0.048)	(0.049)
Believe there is:					
misinformation problem on SM	-0.018	-0.018	-0.019	-0.024	-0.011
	(0.020)	(0.028)	(0.013)	(0.029)	(0.030)
a banning problem on SM	0.032	0.028	0.040^{***}	0.021	0.035
	(0.022)	(0.030)	(0.014)	(0.031)	(0.033)
shadow-banning problem on SM	-0.044^{**}	-0.038	-0.056^{***}	-0.039	-0.037
	(0.022)	(0.031)	(0.014)	(0.032)	(0.033)
an onine bullying problem on SM	0.039^{*}	0.052^{*}	0.012	0.072^{**}	0.032
	(0.022)	(0.031)	(0.014)	(0.032)	(0.033)
a civility problem on SM	0.047^{**}	0.083^{***}	-0.026^{*}	0.091^{***}	0.076^{**}
	(0.021)	(0.030)	(0.014)	(0.031)	(0.032)
Frequently talk to others about politics	0.042	0.047	0.033	0.045	0.048
	(0.038)	(0.053)	(0.025)	(0.055)	(0.057)
Infrequently talk to others about politics	0.034	0.049	0.004	0.067	0.032
	(0.037)	(0.051)	(0.024)	(0.053)	(0.055)
Constant	1.307^{***}	1.718^{***}	0.486^{***}	1.743^{***}	1.692^{***}
	(0.130)	(0.180)	(0.085)	(0.188)	(0.194)
Observations	1,084	1,084	1,084	1,084	1,084
R^2	0.060	0.058	0.070	0.052	0.055
Adjusted R^2	0.046	0.044	0.057	0.038	0.041
Residual Std. Error (df = 1067)	0.372	0.515	0.241	0.535	0.555
F Statistic (df = 16; 1067)	4.280^{***}	4.102^{***}	5.057^{***}	3.674^{***}	3.911^{***}
Note:					*p<0.1; **p<0.05; ***p<0.01

Supplementary Information 4 Content of posts

In developing the content of the simulated posts for our study, we read through a number of partisan-related posts on social media platforms and news sites like X (formerly known as Twitter), Instagram, Fox News, the New York Times, and the Daily Mail.

We were careful not to write content that would include any specific mention of political issues (e.g. about climate change, immigration, health care etc.), because of the potential risk that such content could elicit strong biases based on personal beliefs. Instead, we chose to focus on posts that either praised or criticized political parties (i.e. either via insulting or praising parties and party members, specifically). This way, we could isolate the effect of partisan (negative, positive, or neutral) sentiment from issue-based sentiment, thereby allowing us to better measure reactions to toxic versus non-toxic content without the confounding influence of individual issue positions.

Furthermore, we defined "political" content as any post that explicitly mentioned or referred to political parties, their members, or their policies in a manner that could be perceived as supportive or critical.

SI 4.1 Political posts

- **In-party negative 1**: These dumbass {INPARTY} have done nothing to fix our economy. completely useless.
- **Out-party negative 2**: If you vote {OUTPARTY} you are an idiot. I'm sorry, I don't make the rules.
- **In-party negative 3**: {INPARTY} don't know when you stop. they will keep going lower and lower and eventually they will alienate just about every group.
- **Out-party negative 4**: If anyone ever votes {OUTPARTY}, after all this, they are just blind, brainwashed, and stupid.
- **Out-party positive 1**: {OUTPARTY} are putting the american people first. Thanks to them, we're finally bringing American manufacturing back home and investing in our national infrastucture.

- **In-party positive 2**: Elections are about choice, so make the right one. {INPARTY} have historically creating record job growth & passed so many important bills.

- **In-party positive 3**: The only patriots in america are those who voted for {INPARTY} leaders.

- **Out-party positive 4**: Any {OUTPARTY} elected is a win for democracy.

SI 4.2 Non-partisan posts

- **Neutral 1**: I'm reading this book on conversations between regular republicans and democrats…just normal folks having real conversations. It's been eye opening. We really shouldn't judge people before we get to know them.

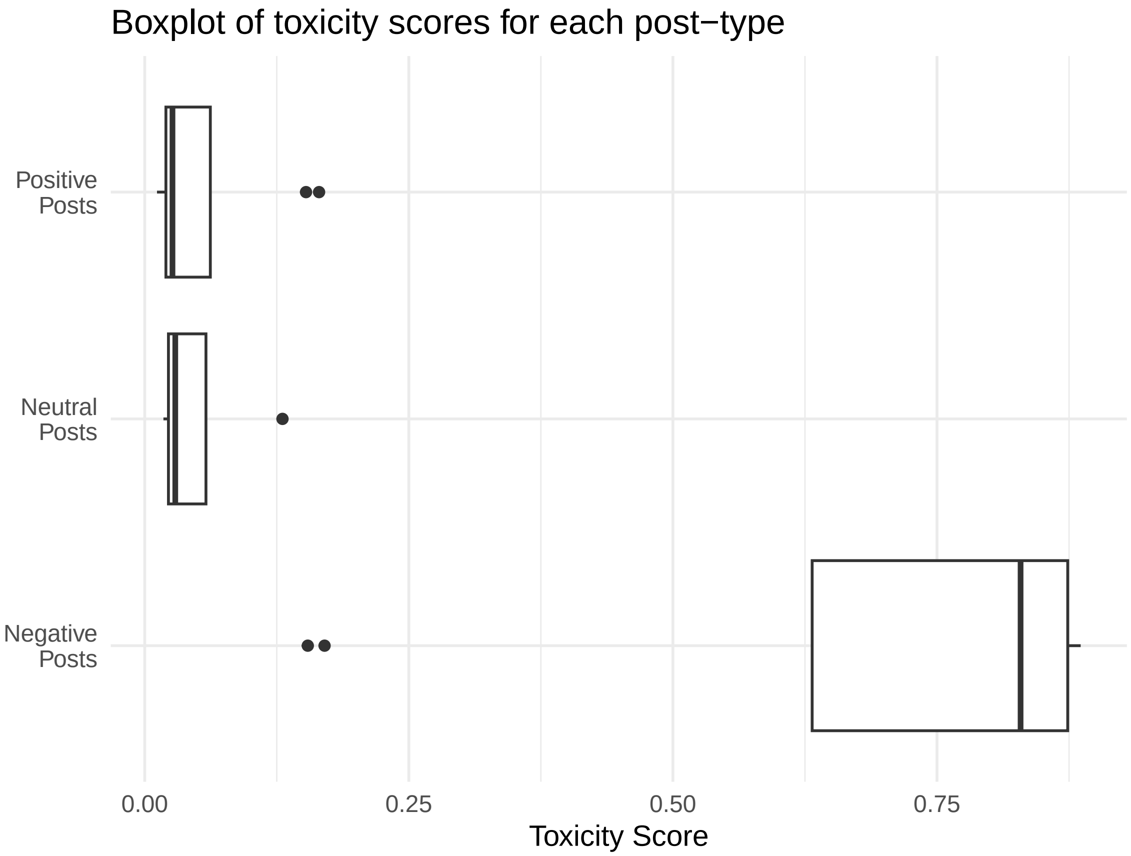
- **Neutral 2**: I hope we can all agree on one thing: there's more to life than just politics. It's really brought out the worst in society.

- **Non-partisan 1**: First day of vacation has officially begun! [Photograph of beach]

- **Non-partisan 2**: Does anyone know what this fruit is called? They look like tiny green apples. They're so sour and I remember eating them with salt. [Photograph of sour green plums]

SI 4.3 Toxicty scores for each post type

The plot below (figure SI 4 shows the toxicity scores for each post category (Positive Posts, Negative Posts, and Neutral Posts) as derived from Google's Perspective API. We can see that our negative posts have very high toxicity scores, whereas the positive and neutral posts yield very low toxicity scores.



Note: Toxicity scores are derived from Google's Perspective API.
Google defines 'toxicity' as the likelihood that a comment is perceived
as disrespectful or likely to make someone leave a discussion.

Figure SI 4. Toxicity scores for each post type. This plot shows that negative posts have much higher toxicity scores than neutral and positive posts. Note that these toxicity scores comes from Google's Perspective API who define toxicity as "the likelihood that a comment will be perceived as disrespectful or likely to make someone leave a discussion."

Supplementary Information 5 Regression Tables

SI 5.1 Study 1

SI 5.1.1 Overall effects of choice

Table SI 6. Regression table for Figure 4A. These are the results of the regression models used to estimate the treatment effects of offering choice on participants' evaluation of ConversationCircle. The table presents the coefficients and standard errors for each independent variable in the models predicting participants' satisfaction with ConversationCircle, likelihood of recommending it to a friend, and likelihood of using the platform.

	Dependent variable:		
	Satisfaction with ConversationCircle	Recommending to friend	Using theplatform
Treatment condition	0.122	0.089	0.201^{***}
	(0.074)	(0.076)	(0.074)
Democrat	-0.331^{***}	-0.275^{***}	-0.383^{***}
	(0.093)	(0.096)	(0.094)
College educated	-0.240^{***}	-0.208^{***}	-0.209^{***}
	(0.076)	(0.077)	(0.076)
Strong party ID	0.121	0.124	0.073
	(0.076)	(0.078)	(0.077)
Age	0.006^{**}	0.008^{**}	0.005^{*}
	(0.003)	(0.003)	(0.003)
Female	0.139^{*}	0.113	0.160^{**}
	(0.077)	(0.079)	(0.077)
White	-0.166^{*}	-0.105	-0.183^{*}
	(0.097)	(0.099)	(0.097)
Frequency of SM use	0.154^{***}	0.147^{***}	0.115^{***}
	(0.037)	(0.038)	(0.038)
Satisfied with SM	0.463^{***}	0.513^{***}	0.448^{***}
	(0.082)	(0.084)	(0.083)
Believe there is:			
misinformation problem on SM	-0.052	-0.054	-0.056
	(0.052)	(0.053)	(0.052)
a banning problem on SM	0.263^{***}	0.235^{***}	0.264^{***}
	(0.053)	(0.054)	(0.053)
shadow-banning problem on SM	-0.164^{***}	-0.118^{**}	-0.166^{***}
	(0.052)	(0.053)	(0.052)
an onine bullying problem on SM	0.136^{**}	0.113^{**}	0.128^{**}
	(0.055)	(0.057)	(0.056)
a civility problem on SM	-0.169^{***}	-0.180^{***}	-0.143^{***}
	(0.053)	(0.054)	(0.054)
Frequently talk to others about politics	0.020	-0.007	-0.019
	(0.091)	(0.093)	(0.092)
Infrequently talk to others about politics	-0.514^{***}	-0.574^{***}	-0.520^{***}
	(0.089)	(0.091)	(0.090)
Constant	-1.291^{***}	-1.566^{***}	-1.391^{***}
	(0.332)	(0.339)	(0.333)
Observations	1,597	1,597	1,597
R^2	0.134	0.131	0.126
Adjusted R^2	0.125	0.122	0.118
Residual Std. Error (df = 1580)	1.471	1.503	1.479
F Statistic (df = 16; 1580)	15.238^{***}	14.906^{***}	14.287^{***}
Note:			$^{*}p<0.1$; $^{**}p<0.05$; $^{***}p<0.01$

Table SI 7. Regression table for Figure 4B. These are the results of the regression models used to estimate the treatment effects of offering choice on participants' average hostility ratings of the posts they saw. The table presents the coefficients and standard errors for each independent variable in the models predicting participants' mean hostility ratings for all posts combined, political posts, non-partisan posts, in-party political posts, and out-party political posts.

	Dependent variable:				
	Mean hostility ratings				
	All posts combined	Political posts	Non-partisan posts	In-party political posts	Out-party political posts
Treatment condition	0.030	0.044	0.003	0.040	0.048
	(0.026)	(0.035)	(0.021)	(0.036)	(0.038)
Democrat	0.024	0.037	−0.003	0.013	0.060
	(0.033)	(0.044)	(0.026)	(0.046)	(0.047)
College educated	0.114***	0.141***	0.060***	0.114***	0.168***
	(0.026)	(0.035)	(0.021)	(0.037)	(0.038)
Strong party ID	−0.027	−0.062*	0.044**	−0.044	−0.080**
	(0.027)	(0.036)	(0.021)	(0.037)	(0.039)
Age	−0.002**	−0.001	−0.005***	−0.001	−0.0003
	(0.001)	(0.001)	(0.001)	(0.001)	(0.002)
Female	−0.025	−0.019	−0.036*	0.006	−0.045
	(0.027)	(0.036)	(0.021)	(0.037)	(0.039)
White	−0.007	0.006	−0.034	0.017	−0.005
	(0.034)	(0.045)	(0.027)	(0.047)	(0.049)
Frequency of SM use	−0.009	−0.012	−0.004	−0.015	−0.009
	(0.013)	(0.017)	(0.010)	(0.018)	(0.019)
Satisfied with SM	0.003	0.010	−0.012	0.016	0.005
	(0.029)	(0.039)	(0.023)	(0.040)	(0.042)
Believe there is:					
misinformation problem on SM	−0.009	0.011	−0.049***	−0.002	0.024
	(0.018)	(0.024)	(0.014)	(0.025)	(0.026)
a banning problem on SM	−0.009	−0.032	0.037**	−0.035	−0.029
	(0.019)	(0.025)	(0.015)	(0.026)	(0.027)
shadow-banning problem on SM	0.003	0.030	−0.050***	0.028	0.032
	(0.018)	(0.024)	(0.014)	(0.025)	(0.026)
an onine bullying problem on SM	0.031	0.044*	0.005	0.059**	0.028
	(0.019)	(0.026)	(0.015)	(0.027)	(0.028)
a civility problem on SM	0.031*	0.066***	−0.039***	0.070***	0.063**
	(0.019)	(0.025)	(0.015)	(0.026)	(0.027)
Frequently talk to others about politics	0.090***	0.100**	0.070***	0.097**	0.103**
	(0.032)	(0.043)	(0.025)	(0.044)	(0.046)
Infrequently talk to others about politics	0.057*	0.085**	−0.0002	0.075*	0.094**
	(0.031)	(0.042)	(0.025)	(0.043)	(0.045)
Constant	1.207***	1.508***	0.606***	1.592***	1.424***
	(0.116)	(0.155)	(0.092)	(0.162)	(0.168)
Observations	1,597	1,597	1,597	1,597	1,597
R^2	0.035	0.042	0.070	0.036	0.042
Adjusted R^2	0.025	0.033	0.061	0.026	0.033
Residual Std. Error (df = 1580)	0.515	0.689	0.408	0.717	0.746
F Statistic (df = 16; 1580)	3.595***	4.364***	7.425***	3.656***	4.360***
Note:				*p<0.1; **p<0.05; ***p<0.01	

Estimating probability of opting-in to viewing less toxic content

Table SI 8. Logistic regression table for Figure 5. This is the model used to estimate control participants' probability of opting-in to viewing less toxic political content had they been offered the choice.

	Dependent variable:
	Opt-in to viewing less toxic political content
Social media has a shadow banning problem	-0.310^{***}
	(0.108)
Satisfied with social media	0.442^{***}
	(0.164)
Social media has a online bullying problem	0.196^{*}
	(0.112)
College educated	0.254^{*}
	(0.154)
Frequency of social media Use	0.129
	(0.078)
Democrat	−0.220
	(0.191)
Social media has a banning users problem	−0.102
	(0.106)
White	−0.203
	(0.213)
Social media has a misinformation problem	0.088
	(0.104)
Female	−0.118
	(0.158)
Strong Party ID	−0.103
	(0.156)
Social media has a civility problem	−0.035
	(0.108)
Age	−0.002
	(0.006)
Frequently discusses politics with others	0.155
	(0.187)
Infrequently discusses politics with others	−0.003
	(0.180)
Constant	−0.117
	(0.692)
Observations	817
Log Likelihood	−511.801
Akaike Inf. Crit.	1,055.603
Note:	*p<0.1; **p<0.05; ***p<0.01

Effect of choice on those who opt-in

Table SI 9. Regression table for Figure 6A. These are the results of the regression models used to estimate the treatment effects of offering choice to participants who would have opted-in to viewing less toxic content on participants' evaluation of ConversationCircle. The table presents the coefficients and standard errors for each independent variable in the models predicting participants' satisfaction with ConversationCircle, likelihood of recommending it to a friend, and likelihood of using the platform.

	Dependent variable:		
	Satisfaction with ConversationCircle	Recommending to friend	Using theplatform
Treatment condition	0.166**	0.124	0.284***
	(0.083)	(0.084)	(0.084)
Democrat	−0.322***	−0.284***	−0.354***
	(0.103)	(0.105)	(0.104)
College educated	−0.260***	−0.218**	−0.246***
	(0.086)	(0.087)	(0.087)
Strong party ID	0.014	0.049	−0.023
	(0.085)	(0.087)	(0.086)
Age	0.005	0.007**	0.003
	(0.003)	(0.003)	(0.003)
Female	0.069	0.075	0.100
	(0.086)	(0.088)	(0.087)
White	−0.158	−0.130	−0.178*
	(0.106)	(0.108)	(0.107)
Frequency of SM use	0.195***	0.152***	0.150***
	(0.045)	(0.046)	(0.045)
Satisfied with SM	0.545***	0.609***	0.490***
	(0.095)	(0.097)	(0.096)
Believe there is:			
misinformation problem on SM	−0.074	−0.067	−0.097
	(0.061)	(0.062)	(0.061)
a banning problem on SM	0.338***	0.302***	0.326***
	(0.060)	(0.061)	(0.061)
shadow-banning problem on SM	−0.204***	−0.151**	−0.201***
	(0.058)	(0.059)	(0.059)
an onine bullying problem on SM	0.127**	0.082	0.148**
	(0.063)	(0.065)	(0.064)
a civility problem on SM	−0.196***	−0.178***	−0.180***
	(0.061)	(0.062)	(0.061)
Frequently talk to others about politics	0.050	0.044	−0.003
	(0.101)	(0.104)	(0.103)
Infrequently talk to others about politics	−0.524***	−0.595***	−0.567***
	(0.100)	(0.102)	(0.101)
Constant	−1.368***	−1.478***	−1.348***
	(0.390)	(0.397)	(0.394)
Observations	1,291	1,291	1,291
R^2	0.164	0.156	0.152
Adjusted R^2	0.153	0.145	0.142
Residual Std. Error (df = 1274)	1.301	1.327	1.315
F Statistic (df = 16; 1274)	15.572***	14.680***	14.326***
Note:			*p<0.1; **p<0.05; ***p<0.01

Table SI 10. Regression table for Figure 6B. These are the results of the regression models used to estimate the treatment effects of offering choice to participants who would have opted-in to viewing less toxic content on participants' average hostility ratings of the posts they saw. The table presents the coefficients and standard errors for each independent variable in the models predicting participants' mean hostility ratings for all posts combined, political posts, non-partisan posts, in-party political posts, and out-party political posts.

	Dependent variable:				
	Mean hostility ratings				
	All posts combined	Political posts	Non-partisan posts	In-party political posts	Out-party political posts
Treatment condition	0.078***	0.093**	0.047*	0.085**	0.101**
	(0.029)	(0.038)	(0.025)	(0.040)	(0.041)
Democrat	−0.007	−0.003	−0.013	−0.018	0.011
	(0.036)	(0.047)	(0.031)	(0.049)	(0.051)
College educated	0.125***	0.143***	0.089***	0.117***	0.169***
	(0.030)	(0.039)	(0.026)	(0.041)	(0.043)
Strong party ID	−0.015	−0.056	0.065**	−0.044	−0.068
	(0.030)	(0.039)	(0.025)	(0.041)	(0.042)
Age	−0.002*	−0.0003	−0.006***	−0.001	0.001
	(0.001)	(0.002)	(0.001)	(0.002)	(0.002)
Female	−0.007	0.005	−0.031	0.035	−0.025
	(0.030)	(0.040)	(0.026)	(0.041)	(0.043)
White	0.039	0.068	−0.018	0.081	0.055
	(0.037)	(0.048)	(0.032)	(0.051)	(0.053)
Frequency of SM use	−0.009	−0.008	−0.011	−0.010	−0.006
	(0.016)	(0.021)	(0.013)	(0.021)	(0.022)
Satisfied with SM	−0.030	−0.040	−0.010	−0.038	−0.043
	(0.033)	(0.044)	(0.028)	(0.045)	(0.047)
Believe there is:					
misinformation problem on SM	−0.006	0.028	−0.074***	0.009	0.047
	(0.021)	(0.028)	(0.018)	(0.029)	(0.030)
a banning problem on SM	−0.024	−0.057**	0.041**	−0.058**	−0.056*
	(0.021)	(0.028)	(0.018)	(0.029)	(0.030)
shadow-banning problem on SM	0.037*	0.075***	−0.040**	0.074***	0.077***
	(0.020)	(0.027)	(0.017)	(0.028)	(0.029)
an onine bullying problem on SM	0.020	0.032	−0.004	0.046	0.018
	(0.022)	(0.029)	(0.019)	(0.030)	(0.032)
a civility problem on SM	0.017	0.050*	−0.049***	0.051*	0.049
	(0.021)	(0.028)	(0.018)	(0.029)	(0.030)
Frequently talk to others about politics	0.112***	0.124***	0.089***	0.120**	0.128**
	(0.035)	(0.047)	(0.030)	(0.049)	(0.051)
Infrequently talk to others about politics	0.063*	0.098**	−0.007	0.072	0.124**
	(0.035)	(0.046)	(0.030)	(0.048)	(0.050)
Constant	1.184***	1.417***	0.719***	1.535***	1.298***
	(0.136)	(0.179)	(0.117)	(0.186)	(0.194)
Observations	1,291	1,291	1,291	1,291	1,291
R^2	0.040	0.051	0.095	0.041	0.052
Adjusted R^2	0.028	0.039	0.083	0.029	0.040
Residual Std. Error (df = 1274)	0.454	0.598	0.390	0.623	0.649
F Statistic (df = 16; 1274)	3.352***	4.245***	8.343***	3.438***	4.338***
Note:				*p<0.1; **p<0.05; ***p<0.01	

SI 5.2 Study 2

SI 5.2.1 Overall effects of choice

Table SI 11. Regression table for Figure SI 1A. These are the results of the regression models used to estimate the treatment effects of offering choice on participants' evaluation of ConversationCircle in our follow-up study (where participants saw no negative content at all). The table presents the coefficients and standard errors for each independent variable in the models predicting participants' satisfaction with ConversationCircle, likelihood of recommending it to a friend, and likelihood of using the platform.

	Dependent variable:		
	Satisfaction with ConversationCircle	Recommending to friend	Using the platform
Treatment condition	0.224***	0.244***	0.221**
	(0.085)	(0.089)	(0.089)
Democrat	−0.074	−0.085	−0.037
	(0.101)	(0.105)	(0.106)
College educated	−0.208**	−0.217**	−0.179*
	(0.087)	(0.091)	(0.091)
Strong party ID	0.203**	0.212**	0.222**
	(0.088)	(0.092)	(0.092)
Age	0.009**	0.011***	0.009**
	(0.004)	(0.004)	(0.004)
Female	0.055	0.029	−0.009
	(0.088)	(0.092)	(0.092)
White	−0.238**	−0.172	−0.087
	(0.119)	(0.124)	(0.124)
Frequency of SM use	0.074*	0.073	0.069
	(0.043)	(0.045)	(0.045)
Satisfied with SM	0.425***	0.497***	0.410***
	(0.094)	(0.098)	(0.099)
Believe there is:			
misinformation problem on SM	−0.183***	−0.150**	−0.241***
	(0.059)	(0.062)	(0.062)
a banning problem on SM	0.111*	0.071	0.126**
	(0.061)	(0.064)	(0.064)
shadow-banning problem on SM	−0.104*	−0.065	−0.106*
	(0.060)	(0.062)	(0.062)
an onine bullying problem on SM	0.101*	0.067	0.150**
	(0.060)	(0.063)	(0.063)
a civility problem on SM	−0.056	−0.059	−0.046
	(0.060)	(0.062)	(0.062)
Frequently talk to others about politics	−0.105	−0.066	−0.061
	(0.107)	(0.112)	(0.112)
Infrequently talk to others about politics	−0.342***	−0.384***	−0.247**
	(0.103)	(0.108)	(0.108)
Constant	−0.251	−0.727*	−0.756*
	(0.381)	(0.398)	(0.399)
Observations	1,086	1,086	1,086
R^2	0.086	0.090	0.076
Adjusted R^2	0.073	0.076	0.062
Residual Std. Error (df = 1069)	1.389	1.451	1.455
F Statistic (df = 16; 1069)	6.320***	6.576***	5.504***
Note:			*p<0.1; **p<0.05; ***p<0.01

Table SI 12. Regression table for Figure SI 1B. These are the results of the regression models used to estimate the treatment effects of offering choice on participants' average hostility ratings of the posts they saw in our follow-up study (where participants saw no negative content at all). The table presents the coefficients and standard errors for each independent variable in the models predicting participants' mean hostility ratings for all posts combined, political posts, non-partisan posts, in-party political posts, and out-party political posts.

	Dependent variable:				
			Mean hostility ratings		
	All posts combined	Political posts	Non-partisan posts	In-party political posts	Out-party political posts
Treatment condition	0.005	0.018	−0.008	0.032	0.004
	(0.037)	(0.052)	(0.035)	(0.054)	(0.062)
Democrat	0.058	0.015	0.102**	−0.057	0.086
	(0.044)	(0.062)	(0.041)	(0.064)	(0.073)
College educated	0.192***	0.230***	0.155***	0.239**	0.221***
	(0.038)	(0.053)	(0.036)	(0.055)	(0.063)
Strong party ID	0.055	−0.041	0.151***	−0.161***	0.079
	(0.039)	(0.054)	(0.036)	(0.055)	(0.064)
Age	−0.005***	−0.004*	−0.006***	−0.007***	−0.0001
	(0.002)	(0.002)	(0.001)	(0.002)	(0.003)
Female	−0.017	−0.028	−0.005	−0.005	−0.051
	(0.039)	(0.054)	(0.036)	(0.056)	(0.064)
White	0.036	0.101	−0.029	0.073	0.130
	(0.052)	(0.073)	(0.049)	(0.075)	(0.086)
Frequency of SM use	−0.019	0.006	−0.045**	−0.015	0.028
	(0.019)	(0.026)	(0.018)	(0.027)	(0.031)
Satisfied with SM	0.088**	0.060	0.115***	0.140**	−0.019
	(0.041)	(0.058)	(0.039)	(0.060)	(0.068)
Believe there is: misinformation problem on SM	−0.089***	−0.033	−0.145***	−0.025	−0.041
	(0.026)	(0.036)	(0.024)	(0.037)	(0.043)
a banning problem on SM	0.012	−0.003	0.026	−0.049	0.043
	(0.027)	(0.037)	(0.025)	(0.039)	(0.044)
shadow-banning problem on SM	0.018	0.045	−0.010	0.019	0.072*
	(0.026)	(0.037)	(0.025)	(0.038)	(0.043)
an onine bullying problem on SM	0.021	0.016	0.025	0.010	0.022
	(0.026)	(0.037)	(0.025)	(0.038)	(0.044)
a civility problem on SM	0.028	0.084**	−0.028	0.093**	0.075*
	(0.026)	(0.036)	(0.024)	(0.038)	(0.043)
Frequently talk to others about politics	0.069	0.014	0.125***	−0.016	0.043
	(0.047)	(0.066)	(0.044)	(0.068)	(0.078)
Infrequently talk to others about politics	−0.013	0.007	−0.032	−0.008	0.021
	(0.045)	(0.063)	(0.042)	(0.065)	(0.075)
Constant	0.843***	0.857***	0.829***	1.345***	0.368
	(0.167)	(0.233)	(0.156)	(0.241)	(0.277)
Observations	1,086	1,086	1,086	1,086	1,086
R^2	0.061	0.034	0.131	0.048	0.035
Adjusted R^2	0.046	0.019	0.118	0.034	0.021
Residual Std. Error (df = 1069)	0.611	0.850	0.571	0.879	1.010
F Statistic (df = 16; 1069)	4.304***	2.346***	10.068***	3.354***	2.426***
Note:					*p<0.1; **p<0.05; ***p<0.01

Estimating probability of opting-in to viewing less toxic content

Table SI 13. This is the logistic regression model used to estimate control participants' probability of opting-in to viewing less toxic political content had they been offered the choice in our follow-up study where participants saw no negative content at all.

	Dependent variable:
	Opt-in to viewing less toxic political content
Social media has a shadow banning problem	-0.231^{*}
	(0.127)
Satisfied with social media	0.137
	(0.202)
Social media has a online bullying problem	-0.127
	(0.132)
College educated	-0.137
	(0.190)
Frequency of social media Use	0.088
	(0.091)
Democrat	0.154
	(0.218)
Social media has a banning users problem	-0.137
	(0.127)
White	-0.038
	(0.263)
Social media has a misinformation problem	0.048
	(0.124)
Female	-0.364^{*}
	(0.191)
Strong Party ID	0.376^{*}
	(0.195)
Social media has a civility problem	0.093
	(0.131)
Age	0.008
	(0.008)
Frequently discusses politics with others	-0.379^{*}
	(0.228)
Infrequently discusses politics with others	0.170
	(0.233)
Constant	0.375
	(0.788)
Observations	556
Log Likelihood	-345.222
Akaike Inf. Crit.	722.444
Note:	*p<0.1; **p<0.05; ***p<0.01

Effect of choice on those who opt-in

Table SI 14. Regression table for Figure 7A. These are the results of the regression models used in our follow-up study (where participants saw no negative content at all), to estimate the treatment effects of offering choice to participants who would have opted-in to viewing less toxic content on participants' evaluation of ConversationCircle. The table presents the coefficients and standard errors for each independent variable in the models predicting participants' satisfaction with ConversationCircle, likelihood of recommending it to a friend, and likelihood of using the platform.

	Dependent variable:		
	Satisfaction with ConversationCircle	Recommending to friend	Using theplatform
Treatment condition	0.270***	0.298***	0.300***
	(0.094)	(0.098)	(0.099)
Democrat	−0.132	−0.129	−0.129
	(0.110)	(0.114)	(0.115)
College educated	−0.202**	−0.269***	−0.159
	(0.098)	(0.101)	(0.102)
Strong party ID	0.191*	0.217**	0.200*
	(0.098)	(0.102)	(0.102)
Age	0.011***	0.012***	0.012***
	(0.004)	(0.004)	(0.004)
Female	0.036	−0.004	0.015
	(0.097)	(0.101)	(0.102)
White	−0.335***	−0.215	−0.182
	(0.129)	(0.134)	(0.135)
Frequency of SM use	0.057	0.056	0.067
	(0.048)	(0.050)	(0.050)
Satisfied with SM	0.469***	0.531***	0.365***
	(0.106)	(0.111)	(0.112)
Believe there is:			
misinformation problem on SM	−0.258***	−0.234***	−0.282***
	(0.067)	(0.070)	(0.071)
a banning problem on SM	0.095	0.051	0.106
	(0.066)	(0.068)	(0.069)
shadow-banning problem on SM	−0.122*	−0.092	−0.130*
	(0.063)	(0.066)	(0.067)
an onine bullying problem on SM	0.124*	0.115*	0.176**
	(0.067)	(0.070)	(0.070)
a civility problem on SM	−0.043	−0.064	−0.063
	(0.066)	(0.068)	(0.069)
Frequently talk to others about politics	−0.123	−0.072	−0.060
	(0.122)	(0.127)	(0.128)
Infrequently talk to others about politics	−0.360***	−0.347***	−0.292**
	(0.112)	(0.116)	(0.117)
Constant	0.041	−0.422	−0.556
	(0.420)	(0.437)	(0.441)
Observations	888	888	888
R^2	0.110	0.108	0.091
Adjusted R^2	0.094	0.092	0.075
Residual Std. Error (df = 871)	1.235	1.284	1.295
F Statistic (df = 16; 871)	6.724***	6.585***	5.469***
Note:			*p<0.1; **p<0.05; ***p<0.01

Table SI 15. Regression table for Figure 7B. These are the results of the regression models used in our follow-up study (where participants saw no negative content at all), to estimate the treatment effects of offering choice to participants who would have opted-in to viewing less toxic content on participants' average hostility ratings of the posts they saw. The table presents the coefficients and standard errors for each independent variable in the models predicting participants' mean hostility ratings for all posts combined, political posts, non-partisan posts, in-party political posts, and out-party political posts.

	Dependent variable:				
	Mean hostility ratings				
	All posts combined	Political posts	Non-partisan posts	In-party political posts	Out-party political posts
Treatment condition	0.082*	0.111*	0.054	0.101*	0.120*
	(0.043)	(0.059)	(0.041)	(0.061)	(0.069)
Democrat	0.047	−0.001	0.094**	−0.082	0.081
	(0.050)	(0.069)	(0.047)	(0.071)	(0.081)
College educated	0.206***	0.230***	0.182***	0.239***	0.221***
	(0.044)	(0.061)	(0.042)	(0.063)	(0.071)
Strong party ID	0.070	−0.022	0.162***	−0.150**	0.105
	(0.044)	(0.061)	(0.042)	(0.063)	(0.071)
Age	−0.006***	−0.005*	−0.007***	−0.008***	−0.001
	(0.002)	(0.002)	(0.002)	(0.003)	(0.003)
Female	−0.030	−0.044	−0.017	−0.032	−0.055
	(0.044)	(0.061)	(0.042)	(0.063)	(0.071)
White	0.048	0.127	−0.030	0.083	0.172*
	(0.058)	(0.080)	(0.056)	(0.083)	(0.094)
Frequency of SM use	−0.032	−0.006	−0.059***	−0.018	0.007
	(0.022)	(0.030)	(0.021)	(0.031)	(0.035)
Satisfied with SM	0.079	0.041	0.117**	0.108	−0.027
	(0.048)	(0.066)	(0.046)	(0.069)	(0.078)
Believe there is:					
misinformation problem on SM	−0.118***	−0.059	−0.177***	−0.049	−0.070
	(0.030)	(0.042)	(0.029)	(0.044)	(0.049)
a banning problem on SM	0.008	−0.006	0.022	−0.049	0.037
	(0.030)	(0.041)	(0.028)	(0.042)	(0.048)
shadow-banning problem on SM	0.040	0.067*	0.013	0.036	0.098**
	(0.029)	(0.040)	(0.027)	(0.041)	(0.046)
an onine bullying problem on SM	0.011	0.010	0.012	−0.003	0.022
	(0.030)	(0.042)	(0.029)	(0.043)	(0.049)
a civility problem on SM	0.024	0.076*	−0.029	0.087**	0.066
	(0.030)	(0.041)	(0.028)	(0.043)	(0.048)
Frequently talk to others about politics	0.131**	0.085	0.177***	0.049	0.121
	(0.055)	(0.076)	(0.053)	(0.079)	(0.090)
Infrequently talk to others about politics	−0.007	0.033	−0.048	0.012	0.054
	(0.051)	(0.070)	(0.048)	(0.072)	(0.082)
Constant	0.999***	0.999***	0.999***	1.471***	0.527*
	(0.190)	(0.262)	(0.181)	(0.271)	(0.307)
Observations	888	888	888	888	888
R^2	0.092	0.044	0.174	0.048	0.049
Adjusted R^2	0.076	0.027	0.159	0.031	0.031
Residual Std. Error (df = 871)	0.558	0.768	0.533	0.796	0.903
F Statistic (df = 16; 871)	5.536***	2.535***	11.443***	2.766***	2.782***
Note:					*p<0.1; **p<0.05; ***p<0.01

SI 5.3 Comparing treatment effects between study 1 and study 2

Table SI 16. Regression results using the combined dataset described in the Results section of the main paper. Here, we compare the treatment effects of those who opted-into (or would have opted-into) viewing less toxic political content in our original study and the followup study. While we see a statistically significant treatment effect across the studies, the interaction term was effectively zero and not statistically significant.

	Dependent variable:	
	Mean hostility ratings	
	Only positive political posts	Positive and non-partisan posts
Treatment condition	0.127**	0.088**
	(0.056)	(0.036)
Study 2	0.283***	0.199***
	(0.063)	(0.041)
Democrat	0.018	0.029
	(0.052)	(0.034)
College educated	0.236***	0.185***
	(0.044)	(0.029)
Strong party ID	0.087**	0.100***
	(0.044)	(0.029)
Age	−0.005***	−0.006***
	(0.002)	(0.001)
Female	−0.073	−0.050*
	(0.045)	(0.029)
White	0.025	−0.002
	(0.056)	(0.037)
Frequency of SM use	−0.023	−0.028*
	(0.023)	(0.015)
Satisfied with SM	−0.081	−0.013
	(0.049)	(0.032)
Believe there is:		
misinformation problem on SM	−0.042	−0.084***
	(0.031)	(0.020)
a banning problem on SM	0.020	0.028
	(0.031)	(0.020)
shadow-banning problem on SM	0.077***	0.029
	(0.030)	(0.019)
an onine bullying problem on SM	−0.001	0.003
	(0.032)	(0.021)
a civility problem on SM	−0.003	−0.023
	(0.031)	(0.020)
Frequently talk to others about politics	0.133**	0.127***
	(0.054)	(0.035)
Infrequently talk to others about politics	0.097*	0.033
	(0.052)	(0.034)
Treatment condition * Study 2	−0.042	−0.020
	(0.088)	(0.057)
Constant	0.701***	0.757***
	(0.200)	(0.130)
Observations	2,179	2,179
R^2	0.055	0.087
Adjusted R^2	0.047	0.080
Residual Std. Error (df = 2160)	0.882	0.574
F Statistic (df = 18; 2160)	6.990***	11.465***
Note:	*p<0.1; **p<0.05; ***p<0.01	

Supplementary Information 6 MTurk HIT Description

Title: Social media and online political discussions

Short Description: We would like to know your opinions about social media and online political discussions to help develop better social media platforms.

Detailed Description: We would like to know your opinions about social media and political discussions. Complete a quick survey for $1.82.

Click the link below to be taken to the survey. When the HIT has been completed, you will receive a code to paste into the box below to receive credit for your participation.

Make sure to leave this window open as you complete the survey. When you are finished, you will return to this page to paste the code into the box.

Supplementary Information 7 Preregistration

SI 7.1 Original study

This is an anonymized version of the pre-registration. It was created by the author(s) to use during peer-review. A non-anonymized version (containing author names) will be made available by the authors when the work it supports is made public.

1) Have any data been collected for this study already? No, no data have been collected for this study yet.

2) What's the main question being asked or hypothesis being tested in this study?

If given a choice, is it likely that users would willingly choose to see content that has been algorithmically selected for them?

Would user choice affect evaluation of polarizing content?

Would user choice affect overall evaluation of social media platforms?

Would allowing users the agency to control the type of content they consume lead to an increase, decrease, or no change in affective polarization (i.e., would they view in- and out-partisans more or less favorably)?

Would allowing users the agency to control the type of content they consume lead them to approaching disagreements with more understanding and empathy?

3) Describe the key dependent variable(s) specifying how they will be measured.

Our main dependent variables are participants' evaluations of social media posts and the "new" social media platform. Participants will be asked to evaluate 12 different posts on how engaging and hostile they are. The ordinal scale for these evaluations are: Extremely, Very, Moderately, Slightly, Not at all. Participants will also be asked how likely or unlikely they would use this new social media platform, whether they felt that the platform favored Democrats, Republicans, or neither, and, finally, how civil or uncivil the posts they rated were.

The 12 social media posts will consist of: two non-political post, two positive in-partisan posts, two positive out-partisan posts, two negative in-partisan posts, two negative out-partisan posts, and two neutral partisan posts. Of the partisan posts, participants from both parties will be shown the same text, however, the party mentioned in the post will be flipped for Democrats and Participants. For example, Democrats will see a post that says "If you vote Democrat, you're an idiot. I'm sorry, I don't make the rules." Republicans will see the same post, with "Republicans" mentioned instead of "Democrats."

Our second main dependent variable of interest is affective polarization. After users evaluate posts, they are asked to rate their feelings towards Republicans and Democrats using thermometer ratings.

4) How many and which conditions will participants be assigned to?

Participants will be randomized into one of two experimental conditions. In both the treatment and control condition, participants will be told that “we want [their] feedback on ConversationCircle, a new social media platform under development. [They] will be asked to rate 12 posts selected by the algorithm on this platform. For privacy, the usernames and photos will be blurred.”

Those in the control condition will then be taken directly to the evaluation portion of the experiment, where they will be asked to evaluate social media posts as described in question 3, above.

Participants in the treatment condition, however, will also be told that “this new social media platform, called ConversationCircle, is experimenting with giving users the choice to reduce political divisiveness in their news feed. If a user chooses, ConversationCircle’s algorithm will show fewer politically toxic posts and will instead prioritize positive posts. For this evaluation, you can also make the choice to use this algorithm to select the sample of posts that you will evaluate.” They will then be asked if they would like to “rate a sample of posts that have been selected by ConversationCircle’s algorithm?” They can answer yes or no, before they are then shown the 12 social media posts to evaluate.

Of note, there is no algorithm selecting the posts. Participants will be shown 12 of the same social media posts as described in a previous section.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.

There are lots of important statistical questions to consider when analyzing these newly collected data, including how large each of the response categories is, whether there is missing data, and how extreme response selections are.

Addressing each of these concerns will help determine the statistical tools needed in order to analyze the data. For example, if there are missing data and they can be assessed to be missing at random, then they can be imputed using simple techniques like MICE; if they are not missing at random then a careful consideration of the missingness mechanism before analysis or imputation happen; and if there are no missing data then no imputation is needed. Similarly, if individuals select only levels 1 and 5 of a variable it might be more reasonable to dichotomize and study it as a binary outcome, while if they do not it might make more sense to create an aggregate index or study them individually using ordinal regression methods. Bayesian modeling tools can also be used to address these questions and to help study potential heterogeneity in treatment effects.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.

We will exclude participants who don't complete the survey, and those who complete the survey in less than 4 minutes.

7) How many observations will be collected or what will determine sample size?

No need to justify decision, but be precise about exactly how the number will be determined. We are aiming to have an equal number of Democrat and Republican participants, with an aim for 1200 to 2000 participants.

8) Anything else you would like to pre-register?

Pre-treatment, participants will be asked for their demographic information, political ideology, party affiliation, their opinions about social media companies, their thoughts on the role of social media in political polarization, and their knowledge about algorithms.

SI 7.2 Follow-up study

1) Have any data been collected for this study already?

It's complicated. We have already collected some data but explain in Question 8 why readers may consider this a valid pre-registration nevertheless.

2) What's the main question being asked or hypothesis being tested in this study?

If given a choice, is it likely that users would willingly choose to see content that has been algorithmically selected for them?

Would user choice affect evaluation of polarizing content?

Would user choice affect overall evaluation of social media platforms?

Would allowing users the agency to control the type of content they consume lead to an increase, decrease, or no change in affective polarization (i.e., would they view in- and out-partisans more or less favorably)?

This new preregistration is extension of [the original study]. In that study, we found evidence of a backfire effect. Namely, that giving users choice led to them evaluating political posts more negatively than those in control. In this study, we would also like to investigate whether treated participants evaluated posts more critically because they were disappointed that the algorithm didn't deliver on its promises (that it would show them less toxic posts). To do this, we ask: Will treated users evaluate positive political posts as more hostile than participants in control?

3) Describe the key dependent variable(s) specifying how they will be measured.

Our main dependent variables are participants' evaluations of social media posts and the "new" social media platform. Participants will be asked to evaluate 8 different posts on how engaging and hostile they are. The ordinal scale for these evaluations are: Extremely, Very, Moderately, Slightly, Not at all. Participants will also be asked how likely or unlikely they would use this new social media platform, whether they felt that the platform favored Democrats, Republicans, or neither, and, finally, how civil or uncivil the posts they rated were.

The 8 social media posts will consist of: two non-political post, two positive in-partisan posts, two positive out-partisan posts, and two neutral partisan posts. Of the partisan posts, participants from both parties will be shown the same text, however, the party mentioned in the post will be flipped for Democrats and Participants. For example, Democrats will see a post that says "If you vote Democrat, you're an idiot. I'm sorry, I don't make the rules." Republicans will see the same post, with "Republicans" mentioned instead of "Democrats."

Our second main dependent variable of interest is affective polarization. After users evaluate posts, they are asked to rate their feelings towards Republicans and Democrats using thermometer ratings.

4) How many and which conditions will participants be assigned to?

Participants will be randomized into one of two experimental conditions. In both the treatment and control condition, participants will be told that “we want [their] feedback on ConversationCircle, a new social media platform under development. [They] will be asked to rate 8 posts selected by the algorithm on this platform. For privacy, the usernames and photos will be blurred.”

Those in the control condition will then be taken directly to the evaluation portion of the experiment, where they will be asked to evaluate social media posts as described in question 3, above.

Participants in the treatment condition, however, will also be told that “this new social media platform, called ConversationCircle, is experimenting with giving users the choice to reduce political divisiveness in their news feed. If a user chooses, ConversationCircle’s algorithm will show fewer politically toxic posts and will instead prioritize positive posts. For this evaluation, you can also make the choice to use this algorithm to select the sample of posts that you will evaluate.” They will then be asked if they would like to “rate a sample of posts that have been selected by ConversationCircle’s algorithm?” They can answer yes or no, before they are then shown the 8 social media posts to evaluate.

Of note, there is no algorithm selecting the posts. Participants will be shown 8 of the same social media posts as described in a previous section.

5) Specify exactly which analyses you will conduct to examine the main question/hypothesis.

There are lots of important statistical questions to consider when analyzing these newly collected data, including how large each of the response categories is, whether there is missing data, and how extreme response selections are.

Addressing each of these concerns will help determine the statistical tools needed in order to analyze the data. For example, if there are missing data and they can be assessed to be missing at random, then they can be imputed using simple techniques like MICE; if they are not missing at random then a careful consideration of the missingness mechanism before analysis or imputation happen; and if there are no missing data then no imputation is needed. Similarly, if individuals select only levels 1 and 5 of a variable it might be more reasonable to dichotomize and study it as a binary outcome, while if they do not it might make more sense to create an aggregate index or study them individually using ordinal regression methods. Bayesian modeling tools can also be used to address these questions and to help study potential heterogeneity in treatment effects.

6) Describe exactly how outliers will be defined and handled, and your precise rule(s) for excluding observations.

We will exclude participants who: don't complete the survey, complete the survey in less than 4 minutes, are located outside the United States, or have taken the survey more than once.

7) How many observations will be collected or what will determine sample size?

No need to justify decision, but be precise about exactly how the number will be determined. We are aiming to have a roughly equal number of Democrat and Republican participants, with an aim for 1000 to 2000 participants.

8) Anything else you would like to pre-register?

Pre-treatment, participants will be asked for their demographic information, political ideology, party affiliation, their opinions about social media companies, their thoughts on the role of social media in political polarization, and their knowledge about algorithms.

Regarding Q1: We have collected 100 responses at the time of the pre-registration. But we have not downloaded them, analyzed them, nor have we looked at them at this time.

SI 7.3 Deviation from Preregistration

Supplementary Information 8 Survey Instrument

q1.1 Key Information:

Thank you for your interest in participating in this survey by researchers at Duke University. The purpose of this research is to understand whether and how social media platforms shape political discussions. Your responses will help inform future research into developing better social media platforms.

The survey should not take longer than 10-15 minutes of your time. Your participation in this activity is entirely voluntary, and you have the right to stop participating at any point.

However, you must answer all survey questions in order to be compensated $1.82 for your time. Remember, you must proceed to the final screen of the study in order to receive your completion code which you must submit in order to be paid.

In accordance with Mechanical Turk policies, we may reject your work if the HIT was not completed correctly, if you fail attention checks, or the instructions were not followed.

Your responses will be linked only by a numeric identifier assigned to you by MTurk. We do not ask for your name or any other information that might identify you. Although collected data may be made public or used for future research purposes, your identity will always remain confidential.

For answers to any questions you may have about this survey, please contact [REDACTED] at [REDACTED]@[REDACTED].com. For questions about your rights as a participant contact the Duke Campus Institutional Review Board at campusirb@duke.edu. Please reference Protocol ID #2023-0046 in your email. We encourage you to print or save this form for your own records.

I am a US resident, at least 18 years of age, and desire of my own free will to participate in this study

- Yes

- No

q2.1 What is your year of birth?

q2.2 What best describes you?

- Male
- Female
- Other __________
- Would rather not disclose

q2.3 Are you of Hispanic, Latino, or Spanish origin?

- Yes
- No

q2.4 Which best describes you?

Select all that apply.

- American Indian or Alaskan Native
- Black or African American
- Asian
- Native Hawaiian or Pacific Islander
- White
- Other ________

q2.5 What is the highest level of school you have completed or the highest degree you have received?

- Did not graduate high school
- High school graduate - High school diploma or equivalent (for example: GED)
- Some college but no degree
- Associate degree in college
- Bachelor’s degree (For example: BA, AB, BS)
- Master’s degree (For example: MA, MS, MEng, MEd, MSW, MBA)
- Professional school Degree (For example: MD,DDS,DVM,LLB,JD)
- Doctorate degree

q3.1 When it comes to politics, would you describe yourself as…

- Extremely conservative.
- Conservative.
- Somewhat conservative.
- Moderate; middle of the road.
- Somewhat liberal.
- Liberal.
- Extremely liberal.

q3.2 Generally speaking do you consider yourself a…

- Republican
- Democrat
- Independent
- Something else

[Randomly show either q4.1 or q5.1] q4.1 If you had to choose, do you think of yourself closer to the Republican Party or the Democratic Party?

- Closer to the Republican Party
- Closer to the Democratic Party

q5.1 If you had to choose, do you think of yourself closer to the Republican Party or the Democratic Party?

- Closer to the Democratic Party
- Closer to the Republican Party

q6.1 [If Participant answers Democrat or Republican in q3.2] Would you call yourself a strong [Democrat/Republican] or a not very strong [Democrat/Republican]

- Strong [Democrat/Republican]
- Not very strong [Democrat/Republican]

q7.1 Thank you for your responses! Next, we'd like to ask you some questions about your opinions on social media platforms.

q7.2 Which social media platforms have you visited or used in the past year?

Mark all that apply.

- Facebook
- Twitter
- Instagram
- Reddit
- Youtube
- Snapchat
- TikTok
- Truth Social
- None of these

q7.3 Generally speaking, how often would you say you used any form of social media over the past month?

- Many times every day
- A few times every day
- About once a day
- A few times each week
- About once a week
- Once or twice a month
- Less than once a month
- Never

q7.4 We'd like you to rate how you feel towards social media companies, as a whole, on a scale of 0 to 100.

Ratings between 51 and 100 mean that you feel positive and favorable towards social media companies. Ratings between 0 and 49 mean that you don't feel favorable toward them. A rating of 50 means you don't feel particularly warm or cold towards social media companies.

How would you rate your feeling toward social media companies, as a whole?

- ____

End of Block: Is respondent a social media user?

Display q8.1 Question: If q7.2 != None of these Or q7.3 != Never

q8.1 Overall, how satisfied or dissatisfied are you with your experience using social media platforms?

- Completely satisfied
- Mostly satisfied
- Somewhat satisfied
- Neither satisfied or dissatisfied
- Somewhat dissatisfied
- Mostly dissatisfied
- Completely dissatisfied

q8.2 How much of a problem, if at all, do you think each of the following is on social media?

q8.2.1 The tone or civility of discussions:

- A major problem,

- A significant problem,
- A minor problem,
- Little to no problem at all

q8.2.2 Companies banning users from their platforms

- A major problem,
- A significant problem,
- A minor problem,
- Little to no problem at all

q8.2.3 Companies limiting the visibility of certain posts

- A major problem,
- A significant problem,
- A minor problem,
- Little to no problem at all

q8.2.4 Harassment or abuse from other users

- A major problem,
- A significant problem,
- A minor problem,
- Little to no problem at all

q8.2.5 Inaccurate or misleading information

- A major problem,
- A significant problem,
- A minor problem,
- Little to no problem at all

q9.1 How much of a responsibility, if any at all, do social media companies have to reduce politically divisive content?

- A major responsibility
- A significant responsibility
- A minor responsibility
- Little to no responsibility

q9.2 How much of an impact, if any at all, do you think social media companies have had on increasing political division?

- A major impact
- A significant impact
- A minor impact
- Little to no impact

q10.1 Based on what you know, which of the following best defines an algorithm?

- A process or set of rules to be followed to perform tasks
- A set of universal laws that govern the creation of all computer software
- A piece of hardware in all computers used to perform calculations on complex datasets
- An archive that provides details about all websites on the internet
- I don't know

q10.2 Computer technology can be trained to review large amounts of information and learn to identify patterns. This technology, called **algorithms (or sets of rules)**, are widely used by social media companies to deliver content that is deemed more engaging for the user.

For example, the posts which are recommended to users as they scroll through their newsfeeds are determined by such algorithms. If the algorithm thinks a user will find something more engaging, it will be shown first.

q10.3 Social media platforms use algorithms (or sets of rules) that determine which content appears in users' feeds. Would you support or oppose the use of algorithms to reduce divisive political content?

- Strongly support
- Somewhat support
- Somewhat oppose
- Strongly oppose

q11.1 Next, we'd like to ask you a few questions about whether and how you engage in political discussions.

q11.2 How frequently do you discuss politics or government with others, regardless of whether it is in person, over the phone, or online?

- Very frequently

- Somewhat frequently
- Occasionally
- Somewhat infrequently
- Very infrequently

q11.3 Thinking about most other people in the United States, how frequently do you see people engaging in the following behavior when they talk about politics or current events?

q11.4 Saying things they know might not be true in order to prove their point.

- Very frequently
- Somewhat frequently
- Occasionally
- Somewhat infrequently
- Very infrequently

q11.5 Using profanities, calling people names, or mocking them

- Very frequently
- Somewhat frequently
- Occasionally
- Somewhat infrequently
- Very infrequently

Display q11.6: If q3.2 = Democrat Or q4.1 = Closer to the Democratic Party Or q5.1 = Closer to the Democratic Party

q11.6 How frequently do you see Republicans engaging in the following behavior when they talk about politics or current events?

Display q11.7: If q3.2 = Republican Or q4.1 = Closer to the Republican Party Or q5.1 = Closer to the Republican Party

q11.7 How frequently do you see Democrats engaging in the following behavior when they talk about politics or current events?

q11.8 Saying things they know might not be true in order to prove their point.

- Very frequently
- Somewhat frequently
- Occasionally
- Somewhat infrequently
- Very infrequently

q11.9 Using profanities, calling people names, or mocking them

- Very frequently
- Somewhat frequently
- Occasionally
- Somewhat infrequently
- Very infrequently

[**Participants are then randomized into treatment or control conditions.**]

[**If participants is randomized into treatment condition, they will see q12.1 - q12.4**]

q12.1 Next, we want your feedback on **ConversationCircle**, a new social media platform under development. You will be asked to rate 12 posts selected by the algorithm on this platform. For privacy, the usernames and photos will be blurred.

q12.2 This new social media platform, called ConversationCircle, is experimenting with giving users the **choice** to reduce political divisiveness in their news feed. If a user chooses, ConversationCircle algorithm will show fewer politically toxic posts and will instead prioritize positive posts.

For this evaluation, you can also make the choice to use this algorithm to select the sample of posts that you will evaluate.

Do you want to rate a sample of posts that have been selected by ConversationCircle's algorithm?

- Yes, my choice is to evaluate posts that have been **algorithmically-selected** to be less divisive
- No, my choice is to evaluate posts that have been **randomly-selected**

Display q12.3: If q12.2 = Yes, my choice is to evaluate posts that have been **algorithmically-selected** to be less divisive

q12.3 To confirm, your choice is to rate a sample of posts that have been algorithmically-selected.

- Yes, I would like to see posts that have been **algorithmically-selected**
- No, I would like to see posts that have been **randomly-selected**

Display q12.4: If q12.2 = No, my choice is to evaluate posts that have been **randomly-selected**

q12.4 To confirm, your choice is to rate a sample of posts that have been **randomly-selected**.

- Yes, I would like to see posts that have been **randomly-selected**

- No, I would like to see posts that have been **algorithmically-selected**

[**If participants is randomized into control condition, they will see q13.1**]

q13.1 Next, we want your feedback on ConversationCircle, a new social media platform under development. You will be asked to rate 10 posts selected by the algorithm on this platform. For privacy, the user name and photo will be blurred.

[**All participants are then taken to the evaluation portion of the experiment. They are shown the 12 posts described in appendix B in a randomized order.**]

q14.1 These dumbass {Participant's INPARTY} have done nothing to fix our economy. completely useless.

q14.2 How **engaging** is this post, if at all?

- Extremely
- Very
- Moderately
- Slightly
- Not at all

q14.3 How **hostile** is this post, if at all?

- Extremely
- Very
- Moderately
- Slightly

- Not at all

q15.1 If you vote {Participant’s OUTPARTY} you are an idiot. I’m sorry, I don’t make the rules.

q15.2 How **engaging** is this post, if at all?

- Extremely
- Very
- Moderately
- Slightly
- Not at all

q15.3 How **hostile** is this post, if at all?

- Extremely
- Very
- Moderately
- Slightly
- Not at all

q16.1 {Participant’s INPARTY} don’t know when you stop. they will keep going lower and lower and eventually they will alienate just about every group.

q16.2 How **engaging** is this post, if at all?

- Extremely
- Very

- Moderately
- Slightly
- Not at all

q16.3 How **hostile** is this post, if at all?

- Extremely
- Very
- Moderately
- Slightly
- Not at all

q17.1 If anyone ever votes {Participant's OUTPARTY}, after all this, they are just blind, brainwashed, and stupid.

q17.2 How **engaging** is this post, if at all?

- Extremely
- Very
- Moderately
- Slightly
- Not at all

q17.3 How **hostile** is this post, if at all?

- Extremely

- Very
- Moderately
- Slightly
- Not at all

q18.1 {Participant's OUTPARTY} are putting the american people first. Thanks to them, we're finally bringing American manufacturing back home and investing in our national infrastucture.

q18.2 How **engaging** is this post, if at all?

- Extremely
- Very
- Moderately
- Slightly
- Not at all

q18.3 How **hostile** is this post, if at all?

- Extremely
- Very
- Moderately
- Slightly
- Not at all

q19.1 Elections are about choice, so make the right one. {Participant's INPARTY} have historically creating record job growth & passed so many important bills.

q19.2 How **engaging** is this post, if at all?

- Extremely
- Very
- Moderately
- Slightly
- Not at all

q19.3 How **hostile** is this post, if at all?

- Extremely
- Very
- Moderately
- Slightly
- Not at all

q20.1 The only patriots in american are those who voted for {Participant's INPARTY} leaders.

q20.2 How **engaging** is this post, if at all?

- Extremely
- Very
- Moderately

- Slightly
- Not at all

q20.3 How **hostile** is this post, if at all?

- Extremely
- Very
- Moderately
- Slightly
- Not at all

q21.1 Any {Participant's OUTPARTY} elected is a win for democracy.

q21.2 How **engaging** is this post, if at all?

- Extremely
- Very
- Moderately
- Slightly
- Not at all

q21.3 How **hostile** is this post, if at all?

- Extremely
- Very

- Moderately

- Slightly

- Not at all

q22.1 I'm reading this book on conversations between regular republicans and democrats. . . just normal folks having real conversations. It's been eye opening. We really shouldn't judge people before we get to know them.

q22.2 How **engaging** is this post, if at all?

- Extremely

- Very

- Moderately

- Slightly

- Not at all

q22.3 How **hostile** is this post, if at all?

- Extremely

- Very

- Moderately

- Slightly

- Not at all

q23.1 I hope we can all agree on one thing: there's more to life than just politics. It's really brought out the worst in society.

q23.2 How **engaging** is this post, if at all?

- Extremely
- Very
- Moderately
- Slightly
- Not at all

q23.3 How **hostile** is this post, if at all?

- Extremely
- Very
- Moderately
- Slightly
- Not at all

q24.1 First day of vacation has officially begun! [Photograph of beach]

q24.2 How **engaging** is this post, if at all?

- Extremely
- Very
- Moderately
- Slightly

- Not at all

q24.3 How **hostile** is this post, if at all?

- Extremely
- Very
- Moderately
- Slightly
- Not at all

q25.1 Does anyone know what this fruit is called? They look like tiny green apples. They're so sour and I remember eating them with salt. [Photograph of sour green plums]

q25.2 How **engaging** is this post, if at all?

- Extremely
- Very
- Moderately
- Slightly
- Not at all

q25.3 How **hostile** is this post, if at all?

- Extremely
- Very

- Moderately
- Slightly
- Not at all

q26.1 Thank you so much for those evaluations! Based on the posts you saw, how likely or unlikely would you use ConversationCircle, the new social media platform?

- Extremely likely
- Very likely
- Somewhat likely
- Neither likely nor unlikely
- Somewhat unlikely
- Very unlikely
- Extremely unlikely

q26.2 Overall, would you say that posts you evaluated were more positive towards Democrats, Republicans, or neither?

- Democrats
- Republicans
- Neither

q26.3 Overall, how would you rate the civility or uncivility of the posts you evaluated?

Were they…

- Extremely civil
- Very civil
- Somewhat civil
- Neither civil or uncivil
- Somewhat uncivil
- Very uncivil
- Extremely uncivil

q26.4 How likely or unlikely are you to recommend this new platform to your friends or family?

- Extremely likely
- Very likely
- Somewhat likely
- Neither likely nor unlikely
- Somewhat unlikely
- Very unlikely
- Extremely unlikely

q26.5 Thinking back to earlier in the survey– before you began evaluating the social media posts:

Were you given a choice over what types of posts you viewed? In other words, were you asked whether you wanted to see algorithmically-selected posts?

- Yes, I WAS asked

- No, I was NOT asked

Display q26.6: If q26.5 = Yes, I WAS asked

q26.6 Do you remember what you chose?

- I chose to view algorithmically-selected posts

- I chose to view randomly-selected posts

- Not sure

q27.1 To wrap up, we'd like you to rate how you feel towards ordinary people (i.e., Republican and Democrat voters) on a scale of 0 to 100.

Ratings between 51 and 100 mean that you feel favorable and warm towards the group.

Ratings between 0 and 49 mean that you don't feel favorable toward the group. You would rate a group at the 50 mark if you don't feel particularly warm or cold towards them.

How would you rate your feeling toward Republicans and Democrats? Remember we are asking you to rate ordinary people (e.g., voters, not politicians).

- Republican voters ______________

- Democrat voters ______________

q28.1 In your own words, what would you like to see changed on the social media platforms that you use? Any and all comments are welcome!

- ____

q28.2 We’d love to hear your thoughts about this survey.

Do you have any feedback you’d like to share? These can be comments about any aspect of this survey (for example, a specific survey question, the posts you were asked to evaluate etc.).

- ____

q37.1 Thank you for completing our survey.

This is your survey completion code: ______

Copy this value to paste into MTurk.

When you have copied this ID, please click the next button to submit your survey.