CES ifo Working Papers

13018
2026

September 2026

The Influence of the Vocal Few: Evidence from Social Media Comments

Dante Donati, Lena Song

CES ifo

Imprint:

CESifo Working Papers

ISSN 2364-1428 (digital)

Publisher and distributor: Munich Society for the Promotion
of Economic Research - CESifo GmbH

Poschingerstr. 5, 81679 Munich, Germany
Telephone +49 (0)89 2180-2740

Email office@cesifo.de
https://www.cesifo.org

Editor: Clemens Fuest

An electronic version of the paper may be downloaded free of charge
· from the CESifo website: www.ifo.de/en/cesifo/publications/cesifo-working-papers
· from the SSRN website: www.ssrn.com/index.cfm/en/cesifo/
· from the RePEc website: https://ideas.repec.org/s/ces/ceswps.html

The Influence of the Vocal Few: Evidence from Social Media Comments

Dante Donati and Lena Song*

September 2026

Abstract

Online comment sections let a small number of vocal individuals reach far beyond their own networks. We conduct a large-scale field experiment on Facebook that randomizes the presence and stance of comments beneath posts for a racial justice organization, reaching around one million U.S. users. Opposing comments increase reactions, comments, and link clicks by 15–43 percent relative to no comments, whereas supportive comments have little effect. A complementary survey experiment shows that similar opposing comments make attitudes less progressive and reduce donations to the organization. Through a common feature of online platforms, the vocal few can exert outsized influence.

Keywords: Social Media, Comments, User-Generated Content, Social Influence, Field Experiments

JEL codes: C93, D72, D83, D91, J15, L82, L86

*Donati: Columbia Business School. dd3137@gsb.columbia.edu. Song: University of Illinois Urbana-Champaign. lenasong@illinois.edu. We thank Charles Amuzie from Color of Change for support. We thank George Beknazar-Yuzbashev, Luca Braghieri, Yiting Deng, Sarah Eichmeyer, Ruben Enikolopov, Rafael Jiménez-Durán, Gita Johar, Maria Petrova, Andrea Prat, Carlo Schwarz, Andrey Simonov, and Yanwen Wang, and seminar and conference participants at the Virtual Quantitative Marketing Seminar, University of Chicago, University of Alberta, University of California Berkeley, Bass FORMS Conference, MIT Conference on Digital Experimentation, Berlin School of Economics, Carnegie Mellon University, University of Illinois Urbana-Champaign, Marketing Science, Columbia University, New York University, Workshop on Platform Analytics, AOM, Early Career Behavioral Economics Conference, and CESifo Venice Summer Institute: Workshop on Digital Platforms for helpful feedback. We are grateful to Anna Bezhanishvili, Seungwoo Kim, Jungyun Kim, Thomas Lilly, Daniel Merlau, and Navtej Singh for excellent research assistance. The research was approved by the Institutional Review Boards at Columbia University (AAAU8166). This study was registered in the American Economic Association Registry for randomized controlled trials under trial numbers AEARCTR-0013812 and AEARCTR-0017850 (Donati and Song, 2024, 2026). We acknowledge financial support from the Russell Sage Foundation (Grant G-2309-44994), the Digital Future Initiative and the Bernstein Center at Columbia Business School, and the Provost Office at Columbia University. The authors declare no conflict of interest. All errors are the authors' own.

1 Introduction

Digital technologies have changed the structure and scale of social influence. Online platforms, such as social media, have extended social influence well beyond offline networks and expanded people's exposure to others' opinions. Unlike traditional media with primarily one-way broadcasting, a defining feature of social media is horizontal communication that allows users to share information and opinions among peers (Zhuravskaya, Petrova and Enikolopov, 2020).

Comment sections on social media provide a natural setting to study horizontal communication. Comments are publicly visible and widely read: in our survey of U.S. adults, 86 percent of respondents report sometimes, often, or very often reading comments on social media, and more than 50 percent report doing so often or very often. At the same time, comments are often generated by a small and potentially unrepresentative subset of users (Kim and Noh, 2026), mirroring the concentration of content production in the hands of a small minority of accounts (Grinberg et al., 2019). As comments compete for attention with the post, they can shape how others receive it, even when those who comment are not representative. This effect may be further amplified by engagement-maximizing algorithms that make posts with comments more visible. Understanding the causal effects of comments, and of the views they express, is therefore central to evaluating how social media platforms shape discourse and how they should be governed.

This paper presents a conceptual framework motivated by salience theory (e.g., Bordalo, Gennaioli and Shleifer 2022) and studies how comments shape other users' subsequent on-platform engagement and off-platform attitudes and behavior in two complementary pre-registered experiments. The framework predicts that the presence of comments raises attention to the post regardless of stance, but that only opposing comments, being the more salient, raise subsequent engagement. In collaboration with Color of Change, the largest online racial justice organization in the United States, we ran a large-scale field experiment in which approximately 1 million Facebook users were randomly assigned to see posts about racial justice that had identical content but differed in the presence and stance of the comment section. To examine downstream attitudinal and behavioral effects, we complement the on-platform experiment with a survey experiment on Prolific.

We show that comments causally affect user engagement, attitudes, and behavior, with effects varying by stance. The presence of comments increases attention to posts, but opposing comments further increase on-platform engagement while shifting attitudes in a less progressive direction and reducing donations to a progressive cause. These results suggest that the comment section is an integral component of platform design that can shape outcomes well beyond the platform itself, creating a tension between online engagement and broader offline social objectives, a trade-off that engagement-based ranking algorithms may further amplify (Acemoglu, Ozdaglar and Siderius, 2024).

To establish these findings, we begin by describing how users engage with racial justice content and analyzing the discussions that emerge in the comment sections. Racial justice provides a useful setting because it is a domain of intense public debate and highly polarized views (Pew Research Center, 2024), allowing us to study cross-cutting exposure in a context where the stakes are socially

important. To collect comments and reactions—Facebook's one-click responses to a post, such as *like*, *love*, and *angry*—we advertised posts to approximately 135,000 Facebook users covering several dimensions of racial justice, including voting rights, environmental justice, and criminal justice. We show that engagement patterns differ sharply across areas with different ideological compositions. Users in conservative areas commented at higher rates than those in progressive areas, yet they reacted less frequently. Moreover, comments were overwhelmingly negative and more likely to be offensive in conservative areas. These findings suggest that engagement does not always imply endorsement, especially in ideologically opposed communities. This supports concerns that the voices visible on social media are more extreme than the audience they reach (Bail, 2021).

We then leverage these comments from real user discussions to estimate their causal impact on subsequent engagement. Isolating the effects of comments from those of the original posts is empirically challenging: posts that receive many early comments may have inherent characteristics that make them more engaging ex ante, and platform algorithms tend to recommend posts with higher early activity, further increasing their visibility and subsequent engagement. To address this challenge, we build on existing platform features to design a pipeline that manipulates comment visibility and stance as a novel research instrument. Using this pipeline, we provide causal evidence that the presence and stance of pre-existing comments influence the behavior of other users.

We implement this pipeline in a large-scale field experiment on Facebook. To isolate the effect of comment stance from that of the post itself, we *re-marketed* a subset of the posts from the previous phase of the study—that is, we ran a second paid campaign delivering the same posts, each now pre-populated with two comments, to an audience that had not seen them before. This second campaign reached roughly 1 million new users combined into 18 clusters of ZIP codes, across areas with different ideological compositions. Using Meta's A/B testing infrastructure, in each cluster, we randomized participants into four conditions: (1) no comments (control), (2) supportive comments only, (3) opposing comments only, and (4) a mixed condition with one supportive and one opposing comment. We measured actions to expand the post and comment section, user interactions with the post (reactions, comments, and shares), and direct traffic to the organization's website.

We made several design choices to address potential concerns. To minimize violations of the Stable Unit Treatment Value Assumption (SUTVA), a real-time filtering pipeline hid all new comments, minimizing the influence that users exposed to the same post could have on one another. To limit differences in other visible interactions, we equalized the number of shares and balanced the number of reactions across conditions. To address the risk of divergent delivery—the tendency of ad algorithms to learn and serve treatment arms to different user types based on early engagement (Braun and Schwartz, 2025; Eckles, Gordon and Johnson, 2018)—we followed and augmented best practices from recent work (Burtch et al., 2025). Specifically, we split budgets evenly across arms, launched all ads simultaneously, imposed a one-impression cap per user, and optimized for reach rather than engagement. In addition, we ran the campaign over several weeks to achieve audience saturation for each ad, ensuring that nearly all users within a defined area were reached, to further limit algorithmic learning and divergence. We verified balance across gender, age, and delivery

metrics such as impressions and costs.

We show that comment sections significantly influence subsequent user engagement: the opinions of a small, vocal minority can shape the behavior of a larger audience. Comparing the treatment conditions to the control condition in which posts were displayed without any comments, we find that displaying any pre-populated comments increases all subsequent engagement with the post by 0.065 percentage points, a 13 percent rise relative to the baseline ($p < 0.01$). Further dissecting by types of engagement, we show that the presence of comments increases the likelihood of users clicking to expand the post by about 0.05 percentage points on a 0.25 percent baseline. Consistent with the fact that comment stance is only revealed after a user clicks to expand, there are no significant differences between supportive, opposing, or mixed conditions. However, comment stance has pronounced effects on downstream on-platform engagement: ads with opposing comments generate significantly more interactions and link clicks than those with supportive comments ($p < 0.01$). Relative to the control, opposing stances increase comments, reactions, and shares by roughly 43 percent ($p < 0.05$) and raise click-through rates by 0.034 percentage points (a 15 percent increase, $p < 0.01$; advertising costs per click and per interaction fall proportionately).

These effects vary with the political leaning of the area. The effect of the comment section is largest in conservative areas, where the baseline engagement with the posts is lowest. Stance also matters in these areas: opposing comments substantially increase interactions and click-through rates, whereas the corresponding estimates in progressive areas are small and imprecise. This suggests that identity congruence may affect the impact of comments, as an opposing comment contrasts with the post for every viewer but aligns with a conservative viewer's own stance, although with ad-level data we cannot separate this channel from differences in attention. The influence of comments, therefore, depends on both their stance and the composition of the audience.

We make several design choices in our field experiment to preserve ecological validity. First, we examine the effects of comment sections on a large, diverse population of social media users across a wide range of ZIP codes. Second, users were unaware they were part of an experiment and interacted with the posts as they naturally would. Third, we used comments from real users, ensuring that the content shown to new users reflected genuine Facebook user responses rather than researcher-generated or AI-generated text. Together, these choices allow us to directly study real-world online discourse on a socially important and divisive issue.

At the same time, on-platform engagement metrics capture only part of the picture. To assess how comment sections shape beliefs, attitudes, and costly off-platform behavior, we conducted a complementary survey experiment on Prolific, mirroring the design, creatives, and organic comments used in the Facebook experiment. We find that both supportive and opposing comments increase the time spent viewing the posts by up to 43 percent relative to control. However, opposing comments shift perceptions of the organization and related issues in a less progressive direction, and reduce incentivized donations by 7.2 percent, with no significant differences between Republicans and Democrats. In contrast, opposing comments make Democrats angrier, more annoyed, and less curious and reflective about the post and its comments than they do Republicans. Thus, although

opposing comments amplify engagement on the platform and traffic to the website, they weaken support for the organization and its message off the platform. Taken together, the evidence highlights a tension between short-run engagement gains and downstream attitudinal and behavioral consequences. A back-of-the-envelope fundraising analysis further suggests that the net benefit of tolerating opposing comments depends on the quality of the additional traffic they generate.

The results have several implications for content producers, platforms, and policymakers. For advertisers and campaign managers, opposing comments raise engagement, but stricter moderation may better protect the organization's reputation and downstream support. Importantly, the objectives of platforms that maximize engagement may not align with those of firms, nonprofits, or political actors that care about their legitimacy or about policy support. Regulations such as the European Union's Digital Services Act and the United Kingdom's Online Safety Act already place growing responsibility on platforms for the content they host, including comments as well as posts.[1] Our results support including comment sections in these regulations. Without moderation, comment sections from the vocal few can disproportionately shape the attitudes and behavior of the broader audience.

Our paper makes several contributions. First, we show that the opinions of a small and unrepresentative set of individuals shape the on-platform behavior and the off-platform attitudes and behavior of a much larger audience. Our paper builds upon and expands the literature on the drivers and consequences of user-generated content (UGC), which has largely studied product reviews.[2] Social media comments represent a distinct and understudied form of UGC. They can express opinions on important issues, with societal rather than commercial consequences. A related literature shows that visible social signals change how people respond to content. Cues showing which friends liked an ad raise engagement (Bakshy et al., 2012; Tucker, 2014). Our paper shows that even when comments come from selected and unknown users rather than friends or influencers, they can influence the many people who read the post, including those who never comment.

Second, our findings speak to an emerging literature on the divergence between engagement and other outcomes on digital platforms. Beknazar-Yuzbashev, Jiménez-Durán and Stalinski (2024) show theoretically that platforms may have incentives to promote harmful yet engaging content, and Beknazar-Yuzbashev et al. (2025) provide supporting experimental evidence. Relatedly, Germano, Gómez and Sobbrio (2026) show that algorithms placing greater weight on social interaction signals can increase engagement while also increasing misinformation and polarization. We document a trade-off between on-platform engagement and off-platform outcomes that operates through the stance of the comment sections.

Third, we contribute to the growing literature on the economics of social media (Zhuravskaya, Petrova and Enikolopov, 2020; Aridor et al., 2024). A recent wave of field experiments has studied how platform design choices shape user behavior and attitudes, focusing on algorithmic curation of the news feed (Levy, 2021; Nyhan et al., 2023; Guess et al., 2023*a*,*b*; Gauthier et al., 2026). These

[1] Digital Services Act (2022, EU): `https://eur-lex.europa.eu/eli/reg/2022/2065/oj/eng`; Online Safety Act (2023, UK): `https://www.legislation.gov.uk/ukpga/2023/50/contents`.

[2] See Luca (2015) for a review.

studies have advanced our understanding of the role of algorithms, but less is known about how the social context surrounding a post shapes user responses. The comment section is also a different lever from the feed: feed ranking is set by the platform, while the composition of a comment section can be managed by whoever posts, which makes it available to any firm, nonprofit, or campaign.

Finally, we contribute a methodology for experimentally varying the social context around a post in the field. As manipulating comments is challenging without platform cooperation or the installation of software such as browser extensions, past work has largely used them as an outcome rather than a treatment (e.g., Moehring Forthcoming). We address this challenge by developing an experimental pipeline that leverages platform features to randomize comment visibility and stance. The infrastructure outlined in the paper can be adapted to study the impact of comments in other settings.

The paper is organized as follows: Section 2 introduces a conceptual framework of the influence of the comment section, Section 3 presents the study setting, Section 4 describes the generation and analysis of engagement from real users, Section 5 presents causal evidence on the impact of the comment section on online engagement, Section 6 shows its effects on offline attitudes and behavior, and Section 7 concludes.

2 Conceptual Framework

Comment sections are ubiquitous online. On Facebook and YouTube, the two most widely used platforms in the United States, they appear directly beneath the post. On Reddit, they are the primary mode of discussion. Beyond social media, online news outlets such as The New York Times, The Wall Street Journal, and CNN maintain dedicated comment sections at the end of articles. This proximity means that comments are often read alongside the post or article itself. Such discussions can expose readers to a range of viewpoints and, in principle, support productive conversation on divisive issues. They can also devolve into dismissive or even hostile exchanges. In either case, the exchange is not confined to its participants. Comments are delivered to the same audience as the post, including many who read them without responding.

Those who comment are not drawn at random from the audience. To conceptualize the influence of a comment section dominated by the vocal few, we use a stylized framework motivated by salience theory (Bordalo, Gennaioli and Shleifer, 2012, 2013, 2018, 2022). The salience of a comment section is captured by three features: its prominence before any of its content is visible, how its content contrasts with the post, and how surprising that content is given what the viewer anticipated. Matching our empirical setting in the field experiment, we model a viewer who first decides whether to open the comment section and, if she does, observes the comments alongside the post. She then decides whether to engage, and her attitudes update in response to what she has seen. In line with the experimental design, the framework takes the composition of the comment section as given and does not model who chooses to comment. Appendix A provides details of the setup and derivations. There are two main predictions:

Prediction 1 (Presence and stance act on different margins of engagement)**.** The comment counter makes the post more prominent, and because stance is not revealed until the post is expanded, the presence of comments raises post expansions regardless of stance. Once the comment section is opened, the opposing section is more salient, so it increases interactions and link clicks relative both to a post without comments and to one with a supportive section.

Prediction 2 (Opposing comments move attitudes away from the post)**.** Because comments shift attention away from the post and toward the comment section, an opposing section can pull attitudes toward its own position and against that of the post and, if willingness to donate is increasing in support for the organization, lower donations. A section that repeats the post's position has little effect, since averaging the post with a section that restates it leaves the position unchanged.

The two predictions highlight a trade-off between engagement and downstream objectives. Greater salience of the comment section raises the total attention people pay to the post and its comments, which increases the probability of engagement. At the same time, salience shifts the allocation of that attention from the post toward the comments, which increases the weight placed on the comments relative to the post when attitudes are formed. When the comments oppose the post, the first effect produces more interactions and clicks, while the second pulls attitudes away from the post. Supportive comments, which are less salient and restate the post, raise post expansions like any comment section but have little effect on interactions, clicks, or attitudes.

Salience responds to how much the comments surprise the viewer and contrast with the post, not to how representative they are. Content produced by a vocal few can therefore influence a much larger audience even when it is not representative. We focus on salience as the channel through which comments affect engagement, and in Section 6 we discuss other channels that could affect offline outcomes, such as learning about prevailing norms.

3 Background and Setting

3.1 Comment Sections and Racial Justice

We study comment sections in the context of racial justice, one of the most divisive issues in the United States. In 2020, the George Floyd protests sparked a racial reckoning across the country. On social media, discussions about race and racial justice, such as those with the hashtag #blacklivesmatter, surged in 2020 (Anderson et al., 2020). Yet racial attitudes remain divided in the U.S. According to a Pew Research Center survey in 2024, 80 percent of Democrats or Democratic-leaning independents say that White people benefit from advantages in society that Black people do not have, while only 22 percent of Republicans or Republican-leaning independents expressed the same view (Pew Research Center, 2024).

Our study provides evidence on how these divisions manifest in online discourse and how they influence subsequent users' on-platform engagement and off-platform attitudes and behavior. We partner with Color of Change, the largest online racial justice organization in the United States.

As a nonprofit, Color of Change uses digital platforms to mobilize supporters to hold institutions accountable. We designed social media posts in line with Color of Change's brand guidelines to ensure the ecological validity of our study.

3.2 Social Media Advertising as a Research Tool

To reach a large audience, we deliver the posts as sponsored content on Facebook using Meta Ads Manager under *The Public Square*, a research account created for the project. Our design leverages social media advertising for several reasons. First, understanding the role of comment sections on advertisements is inherently important, as social media ads represent a significant share of the trillion-dollar advertising industry. In 2025, global social media ad spend exceeded $275 billion, including about $100 billion in the U.S.[3] While existing literature has studied social media ad effectiveness (e.g., Gordon et al. 2019), the role of user comments in enhancing or diminishing ad impact remains underexplored. Comments on ads allow users to share information and opinions, potentially influencing future viewers, akin to other forms of UGC. Despite significant interest from firms in leveraging UGC and social influence, there is little empirical evidence on the role of social media comments.

Second, using Facebook advertising allows us to access a large and diverse sample while minimizing experimenter demand effects. Recent work (e.g., Donati and Rao 2025; Donati et al. Forthcoming) has explored social media ads as a research tool, as they enable the delivery of content to a broad audience and allow researchers to observe user behavior in the natural context in which the content is usually encountered.

Third, we develop a novel pipeline for comment section manipulation using existing features in the Meta Ads Manager. This enables us to manage comment sections across a large number of posts and systematically collect relevant platform data. Because our approach relies solely on tools already available on the platform, it yields practical insights for social media managers seeking to moderate and manage comment sections, as well as a methodological contribution for researchers seeking to vary social context in an experimental setting at scale.

3.3 Post Design and Pre-testing

To generate the comment sections used in our analysis, we produced ad creatives and copy on five issues identified with our partner organization: voter suppression, environmental justice, criminal justice and police reform, education reform, and technology fairness.[4] The organization reviewed and approved all materials, so the creatives are comparable to content the organization would itself run rather than researcher-produced stimuli. We pre-tested the designs and advertised the best-performing ones to elicit the organic comments that form the basis of our main experiment.

[3] DataReportal, *Digital 2026: Global Overview Report*, `https://datareportal.com/reports/digital-2026-global-overview-report`.

[4] See Appendix C for a detailed description of these issues.

Appendix Figure C1 displays the creatives and their headlines exactly as they would appear to Facebook users. Appendix C reports the pre-tests used to select among them and to assess their performance under alternative delivery configurations.

3.4 Facebook Audience Selection

We reach Facebook users via sponsored content. A key advantage of social media advertising is that it enables both broad distribution and precise control over audience targeting (Aridor et al., 2026).

To examine how responses vary by audience characteristics, we use ZIP codes as a targeting criterion. We group ZIP codes that are similar in political ideology and racial composition into strata, each pooled into an audience group to target. Within each ideology–race category, we construct several independent strata. Homogeneous strata ensure that arms are compared on similar users rather than on systematically different areas, since Meta's A/B tool randomizes within a single audience. It also lets us compare outcomes across areas with different ideological leaning. The ZIP code characteristics were collected from several sources:

- **Meta Audience Estimates**: We use information on audience size provided by Meta, which reports the estimated number of users advertisers could potentially reach over a given period.[5] This data was collected through the Marketing API for each ZIP code on October 15, 2024.
- **Voting Behavior**: To proxy political preferences and ideology at the ZIP code level, we rely on the 2020 voting results. These come at the precinct level and were obtained from The Upshot.[6] We assigned each precinct to its nearest ZIP code according to Euclidean distance of the centroids using GIS software, and then aggregated voting information across all precincts matched with the same ZIP code.
- **Population and Racial Composition**: We use 2020 Census data to obtain information on the total and Black populations in each ZIP code and compute the share of Black residents.[7]

We categorize ZIP codes into three ideology groups based on the Republican vote share: *Blue* (Republican vote below 30 percent), *Swing* (between 45 and 55 percent), and *Red* (above 70 percent). Each ideology group was further divided into two subgroups based on racial composition (low and high share of Black population relative to the group median). Hence, a total of six ideology-race categories were created. Each ZIP code is used only once during our experiments. Because Meta's A/B tool randomizes users within a single audience, this ensures that each user is exposed to the content in a single, well-defined condition.

[5]Meta Business Help Center, "About Estimated Audience Size," `https://www.facebook.com/business/help/1665333080167380?id=176276233019487`.

[6]The Upshot, *Presidential Precinct Map 2020*, `https://github.com/TheUpshot/presidential-precinct-map-2020`.

[7]U.S. Census Bureau, 2020 ZIP Code Tabulation Area shapefiles, `https://www.census.gov/cgi-bin/geo/shapefiles/index.php?year=2020&layergroup=ZIP+Code+Tabulation+Areas`.

4 Generation and Analysis of Engagement

To analyze the impact of comment sections, we begin with a large-scale social media campaign designed to elicit engagement with posts about racial justice. This initial phase generates user comments and reactions in a naturalistic setting, and provides direct evidence on how individuals engage with divisive content across audience types and topics. This engagement forms the basis for subsequent experimental manipulation.[8]

Generating engagement is an important feature of our design for several reasons. First, comments reflect genuine, realistic user behavior, which is crucial for ensuring external validity. Unlike researcher- or AI-generated content, organic comments capture the authentic language, tone, and perspectives that users produce and encounter on social media. Second, these comments are themselves substantively important to study. By targeting specific audiences through Facebook's ad infrastructure, we can link engagement patterns to detailed audience characteristics—such as demographics and ZIP code–level ideology—providing richer insights than are typically available. Third, using organic comments enhances the ethical integrity of the experiment, as users interact with content created by other real users rather than being unknowingly exposed to artificially constructed narratives.

4.1 Methodology

Between January 13 and February 10, 2025, we ran a Facebook ad campaign featuring five banners—one for each issue—that had achieved higher click-through rates (CTR) in pre-tests. Each banner promoted content related to racial justice, covering topics in education, environmental justice, criminal justice and police reform, technology fairness, and voting rights. The campaign was optimized to maximize engagement with the posts—showing ads to users most likely to react, comment, or share—in order to collect authentic interactions that reflected spontaneous responses to important yet divisive content.

To examine variation in engagement across audiences and ad characteristics, we created 30 distinct audience strata. Each stratum consisted of sets of ZIP codes randomly sampled and grouped by ideological similarity and racial composition, with an estimated Facebook audience size of about 800,000 users on average (we excluded ZIP codes used in pre-tests). The 30 strata corresponded to six combinations of political ideology (conservative, moderate, and liberal) and racial composition (above or below the median share of Black residents within each ideological category). For each of the six combinations (e.g., conservative areas with a below-median Black population share), we constructed five audience groups, yielding a total of 30 audience strata. Within each stratum, we used Facebook's native A/B testing tool to randomly allocate users to be potentially exposed to one of the five issue banners. In total, the campaign included 150 posts (30 strata $\times$ 5 topics), reaching approximately 135,000 individuals and generating 12,000 reactions,

[8]The generation and analysis of engagement, as well as the main field and survey experiments, were pre-registered before being fielded. The analysis follows the pre-analysis plans, and Appendix B documents the few deviations.

1,750 unique link clicks, 1,500 direct comments,[9] and 650 shares.

The randomization from A/B testing enables comparisons of engagement patterns across topics while holding the potential audience composition constant. The resulting data allow us to characterize the intensity and nature of engagement and to identify patterns of interaction by political ideology and demographic composition. However, we do not interpret potential differences as causal effects of content. Although audiences are randomly assigned to potential exposure, actual exposure is determined by Facebook's ad-delivery algorithm, which endogenously allocates impressions based on predicted engagement probabilities (Braun and Schwartz, 2025). As a result, the observed engagement patterns reflect the joint influence of both content characteristics and algorithmic delivery.

We focus on engagement outcomes visible to other users: comments and reactions. A *reaction* is a one-click response from a fixed menu of emoji (*like*, *love*, *care*, *laugh*, *wow*, *sad*, *angry*), shown only as an aggregate count. A *comment* is written text posted under the user's name and profile picture. Commenting is thus both more effortful and more public than reacting, which is why the two can move in opposite directions across audiences and why the comment section need not represent the audience that sees the post.

First, we compare overall engagement levels and rates (expressed as a percentage of total reach) across ideological groups, computing confidence intervals using standard errors clustered at the advertisement level (150 posts). Second, we assess the *valence* of these interactions. For reactions, we classify *likes*, *loves*, and *cares* as supportive of the posts.[10] For comments, we use GPT–4 to characterize their political stance, sentiment, informativeness, and offensiveness, and the Google Perspective API to measure toxicity. Because a reply is often hard to interpret by itself, we include a description of the post in the prompt, so each comment is coded in the context of the content it responds to. Appendix I documents NLP-based measures used in the paper, including the response scales and the API settings.

4.2 Engagement Across Locations

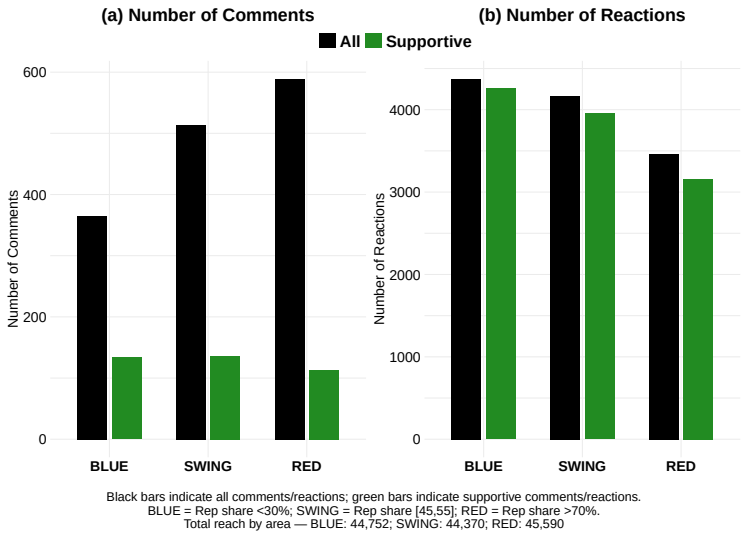
We first document patterns of engagement—both comments and reactions—across different areas. As shown in Figure 1, there are pronounced differences in how users engage with racial justice content across areas with different ideological compositions. When considering all interactions irrespective of their valence, the number of comments increases substantially from liberal to conservative areas. We find a similar pattern for the rate of commenting in Appendix Figure D1. Comment rate rises from about 0.8 percent of reach in Blue ZIP codes to 1.3 percent in Red ones ($p < 0.01$). Reactions, by contrast, follow the opposite pattern, declining from roughly 10 percent of reach in Blue areas to 7.5 percent in Red areas ($p < 0.01$). These differences indicate that users in more conservative areas are less likely to engage through quick, low-effort reactions but more

[9] Direct comments are those directed at the Facebook page/post itself, as opposed to replies to other users' comments.

[10] Other reaction types, such as *laugh*, *wow*, *sad*, and *angry*, are context-dependent and therefore harder to interpret; we focus on those that most reliably convey positive engagement.

likely to participate vocally by commenting on posts. In more progressive areas, engagement occurs primarily through reactions, suggesting a more silent mode of interaction. Because reactions are far more common than comments, summing the two measures across the subfigures indicates that users in progressive areas are, as expected, more likely to engage with racial justice posts overall.[11]

Figure 1: Comment and Reaction Counts by Location



Notes: This figure reports the total number of comments and reactions generated during the initial Facebook campaign, separately by location. Black bars denote all comments/reactions, while green bars denote supportive comments/reactions. Areas are grouped as BLUE (Republican vote share below 30%), SWING (Republican vote share between 45% and 55%), and RED (Republican vote share above 70%). Supportive reactions include likes, loves, and cares.

When focusing on *supportive engagement*—defined as reactions or comments expressing agreement or approval—the ideological gradient is notably flatter for comments. Supportive comments remain consistently low across areas, ranging around 0.3 percent of reach, with no statistically significant differences between Blue, Swing, and Red ZIP codes ($p > 0.20$). In contrast, supportive reactions decline significantly from roughly 9.5 percent in Blue areas to about 7 percent in Red areas ($p < 0.01$), mirroring the overall reaction pattern. This suggests that while the overall volume of vocal participation (comments) rises in conservative areas, supportive responses remain relatively small and stable. Taken together, these results imply that ideological context shapes not only the intensity but also the *type* of engagement: audiences in liberal areas interact more through silent reactions, whereas those in conservative areas engage more vocally, using comments more frequently to express or debate opposing views.

We present the results separately for each issue in Appendix Figure D2 (Appendix Figure D3 reports the rates). Issues such as voter suppression, as well as criminal justice and police reform, generate substantially more vocal engagement in Red ZIP codes, producing large differences relative to Blue ZIP codes. In contrast, topics such as education reform and technology fairness elicit lower levels of vocal engagement overall and exhibit little variation across areas. This pattern suggests

[11] As these are differences across locations, we interpret these results as descriptive of who is exposed to and reacts to racial justice content when it is targeted at an area, not as a causal effect of ideology. Conservative and progressive areas differ on many other characteristics that plausibly affect commenting, such as age composition.

that users in more conservative areas are not uniformly more vocal; rather, specific issues within the broader domain of racial justice appear to trigger heightened vocal engagement and often dissent. Unlike comments, reactions follow the general pattern in which users in more progressive areas are more likely to react, and this pattern holds consistently across all issues.

While the different issues allow us to show that engagement patterns differ even within racial justice, by design all of the posts are progressive. Therefore, we interpret the results as descriptive of how people respond to progressive content, rather than content in general. Nevertheless, the results highlight that the comment section under progressive content is disproportionately from conservative areas, so the vocal few are not representative of the underlying audience.

4.3 Content and Tone of the Comment Section

Appendix Figure D4 shows how the content and tone of comments vary across areas with different ideological compositions, with all rates expressed as a share of total comments.[12] Comments originating from conservative areas are substantially more likely to express a conservative stance and a negative sentiment. The share of conservative-leaning comments rises from about 55 percent in Blue ZIP codes to roughly 75 percent in Red areas (Panel (a), $p < 0.01$), while the share of comments with negative sentiment increases from around 56 percent to nearly 80 percent (Panel (b), $p < 0.01$). Panel (c) shows that the prevalence of offensive language increases from roughly 35 percent in Blue areas to almost 50 percent in Swing and Red areas, with a statistically significant difference relative to Blue areas ($p < 0.01$). Finally, Panel (d) indicates that the share of informative comments—those providing factual content or elaboration—remains relatively low, between 7.5 and 11 percent on average, with no statistically significant differences across locations ($p > 0.10$). We find similar patterns with our pre-specified measures of conversation quality (Appendix Table D2). In the average post, about a third of direct comments engage the existing conversation rather than shifting topic or attacking, fewer than one in five give a reason for the position they take, and fewer than half are civil.

Taken together, these results suggest that while the tone and stance of comments vary, comment sections on divisive issues are generally not very informative across geography.[13] Rather than stating a conservative position with relevant arguments, an opposing comment typically voices disagreement without justification.

5 The Impact of Comments on On-platform Engagement

In this section, we provide causal evidence on the effect of the comment section. Specifically, we study how the presence and stance of a comment section influence subsequent users' on-platform engagement with the content.

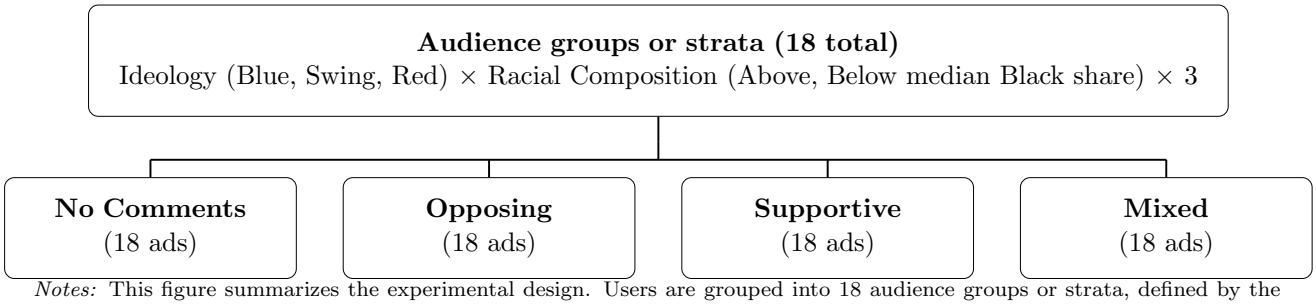
[12] Appendix Tables D1 and D2 report additional pre-specified analyses on the characteristics of the comment sections, including length, depth, and other measures of conversation quality and opinions' diversity.

[13] This could be because commenters simply prefer to express a stance, or because expressive content draws higher engagement than informative content (Lee, Hosanagar and Nair, 2018).

5.1 Design

The experiment ran from March 26 to April 13, 2025, and reached about 1 million Facebook users. Figure 2 provides an overview of the experimental design. We manipulate the comments that participants see below our ads in a new ad campaign, using Meta's A/B testing tool and an automated pipeline that hides new comments from users.

Figure 2: Design Overview



Notes: This figure summarizes the experimental design. Users are grouped into 18 audience groups or strata, defined by the interaction of ZIP code ideology (Blue, Swing, Red) and racial composition (above or below the median Black population share). Within each ZIP code group, users are randomly assigned via Meta's A/B testing infrastructure to one of four comment conditions: (i) no comments (control), (ii) opposing comments only, (iii) supportive comments only, or (iv) mixed comments.

We investigate how exposure to the comment section and the different narratives expressed in the comment section of a post (collected in the Facebook campaign described above) affect individuals' subsequent engagement with that post (views and comments), as well as their intentions (clicks).

5.1.1 Intervention Design

To select and stratify audience types, we follow a similar approach as the one described in Section 4. We exclude ZIP codes used in the comment generation campaign, and create 18 audience groups, with three groups for each combination of political ideology (Blue, Swing, Red) and racial composition (above or below the median share of Black residents within each ideological category).

Within each audience group, we use A/B testing to randomly split its population into four conditions: no visible comments (No Comments), visible comments that include both opposing and supportive comments (Mixed), opposing comments only (Opposing), and supportive comments only (Supportive). Figure 3 provides an example of what users could see under each condition.

Throughout the paper, Supportive and Opposing describe a comment's stance toward the post beneath which it appears, not its position on an absolute political scale. A supportive comment endorses the claim the post advances; an opposing comment rejects it. Because every post in our experiment advocates a progressive position on racial justice, opposing comments in our setting are also comments a reader would code as conservative. We use the relational terms deliberately, since the object we manipulate is the presence of visible agreement or disagreement with the content.

On Facebook, comments on ads are not visible by default. Users must actively click the comment counter or comment button, or expand the ad container, to view the comment section. Importantly,

Figure 3: Example Experimental Conditions



(a) No Comments (b) Supportive (c) Mixed (d) Opposing

Notes: This figure displays screenshots of the four experimental conditions as they appeared to users in their Facebook feed. Panel (a) shows the No Comments condition (control), in which the post is displayed without a visible comment section. Panel (b) shows the supportive condition, in which two supportive comments are pre-populated below the post. Panel (c) shows the Mixed condition, in which one supportive and one opposing comment are displayed. Panel (d) shows the opposing condition, in which two opposing comments are pre-populated below the post. All other features of the post (e.g., the image, caption, call-to-action link, number of shares) are held constant across conditions. Aggregate like and reaction counts vary naturally but are balanced across conditions.

Facebook algorithmically ranks comments based on engagement and other signals.[14] To mitigate potential ranking effects, we display exactly two comments in each comment condition, thereby minimizing within-condition variation in comment visibility. In addition, at the start of the campaign, the share counter is equalized across conditions (12 shares).[15] The comment counter is absent in the control condition and equalized across the comment arms that are part of the same A/B test, displaying 20, 30, or 40 comments. The reaction counter is similarly balanced across posts within the same A/B test, set at approximately 60, 70, or 80. Figure 3 shows an example of an A/B test in which the comment counter is set to 30 in the treatment arms and the reaction counter is approximately 80.

5.1.2 Issue Selection

From the five issues used in the comment generation phase, we selected environmental justice to focus on in order to maximize statistical power. This topic was chosen because it ranks near the middle in terms of overall engagement in the earlier comment-generation campaign, ensuring sufficient variation in user responses without being dominated by extreme levels of attention or controversy. In addition, the topic remained timely and relevant at the time of the experiment.

[14] Meta Newsroom, "Making Public Comments More Meaningful," June 2019, `https://about.fb.com/news/2019/06/making-public-comments-more-meaningful/`.

[15] We equalized the number of shares across posts by sharing them ourselves before the experiment.

5.1.3 Post Selection

We showed a subset of posts from the comment generation campaign with existing interactions to new audience groups. To select the posts for the treatment conditions, we identified triplets of posts using the following procedure. As part of the analysis in Section 4, comments are classified using a 5-point scale for political ideology: Strongly progressive or left-leaning, Slightly or moderately progressive, Centrist/unclear or no explicit stance, Slightly or moderately conservative, or Strongly conservative or right-leaning. In general, comments that support the original post or express pro-racial justice views are classified as progressive, while those that oppose the post or its message are classified as more conservative. To select posts for the Mixed, Opposing, and Supportive conditions, we focused on posts that have at least two supportive and two opposing direct comments (i.e., comments that are directly responding to the post, rather than comments that reply to another comment). For each triplet of posts with a similar number of reactions, we assigned one post each to the Mixed, Opposing, and Supportive conditions.

For the No Comments condition, we selected posts that are not used in any other conditions and have a number of reactions similar to the average number of reactions in each triplet group.

We selected 12 posts, organized into three groups of four. Posts are grouped by reaction count, so posts within a group have similar numbers of reactions. Within each group, we have one post in each of Mixed, Opposing, and Supportive, and No Comments (control). All 12 posts use the same ad banner and headline, with identical image, copy, and call-to-action link. We advertise each post to six distinct audience groups, one in each of the six ideology-race categories, creating 72 advertisements shown to around one million individuals.

5.1.4 Comment Selection

To create the Opposing and Supportive conditions, we selected two comments that matched the relevant stance. For the Mixed condition, we selected one opposing comment and one supportive comment. These comments remained visible to users in the main experiment, while all other direct comments and replies were hidden. We did so using Facebook's existing moderation tool—the *Hide* comment option available to page owners (see Figure E1). This tool makes a comment invisible to other users.

Comments were selected by rule and manual inspection. From each eligible post we kept only direct comments, discarding replies to other users. Among these we gave priority to comments at the two extremes of the five-point stance scale. All displayed comments are coded 1 or 5, except one opposing comment coded 4. Posts within a group were assigned to the Mixed, Opposing, and Supportive conditions starting from a random draw, and we finalized the assignment and the displayed comments by manual inspection, checking that each post had suitable comments for its condition. No displayed comment was edited, shortened, or rewritten, and none was written by the research team. Appendix Table E1 reports the full text of every displayed comment with its GPT-4 scores.

Most of these comments are expressive and often convey a position without justification. Our

comment selection procedure allows us to identify the effect of the stance of a comment section under a progressive post, holding the post content and the audience fixed and other visible interactions balanced.[16]

Finally, to better isolate the effect of the number of comments versus the narrative of the comments on outcomes, we manipulated the number of comments in the posts displayed to audience groups. In the No Comments condition, we deleted as many comments as possible so that the comment counter is zero or close to zero when it is first delivered to the audience in the experiment.[17] In the triplet of posts used for the treatment conditions, we equalized the number of initial comments across conditions by deleting comments or by posting comments ourselves and hiding them, while allowing the total number of comments to vary across triplets (ranging from 20 to 40). This procedure ensured that the comment counter in the No Comments condition was substantially lower than the counter displayed in the treatment conditions.

5.1.5 SUTVA Violations

A potential threat to the internal validity of social media experiments is the violation of the Stable Unit Treatment Value Assumption, or SUTVA (Aridor et al., 2026). This concern is particularly relevant to our research question, given the inherently social nature of comment sections. Because outcomes for one user may depend on the treatment or behavior of others exposed to the same post, interference is a first-order threat, and reducing it through design rather than only through analysis is preferable where feasible. New comments posted by users could influence the perceptions or behaviors of other users, thereby confounding the treatment effects. For example, a new comment supportive of racial justice posted in the opposing condition would alter the intended composition of the comment section.

To address this concern, we implemented a real-time, automated comment-hiding pipeline that immediately hides any new user comments from other users once they are posted.[18] This design ensured that the No Comments condition displayed no comments and that users in the treatment conditions saw only the comments corresponding to the assigned stance. It also helped minimize interactions among users exposed to the same post, thereby mitigating possible SUTVA violations resulting from social interactions within the comment section.[19]

Hiding new comments comes at a cost. In the field, an early comment draws replies and the composition of a section shifts over the life of a post. Existing work shows that what is already

[16] Because every post takes a progressive position, comments that oppose (support) the post are also comments a reader could call conservative (progressive). We therefore cannot separate the effect of disagreement (agreement) with the post from the effect of simply being exposed to conservative (progressive) comments.

[17] Some comments, such as those violating Facebook policy, are not visible to the research team and therefore cannot be deleted. As a result, for some comment sections, it may not be possible to reduce the comment counter to zero.

[18] The commenter will not know that their comment has been hidden; the comment remains visible to the commenter as well as to the page owner.

[19] Other forms of visible engagement on the post, such as the number of likes and shares, remained stable over the treatment period. At the start of the experiment, the average post had 70.92 reactions and 12 shares; the additional engagement generated over the course of the experiment is small relative to these baseline levels.

visible changes what later users write: exposure to higher quality comments leads strangers to write higher quality comments (Berry and Taylor, 2017). Our pipeline shuts down this channel for internal validity: otherwise users exposed to the same post would determine each other's treatment, and the stance we assigned would not be the stance later users saw. We therefore estimate the effect of a fixed comment section on the users who read it, not the joint effect of that section and the discussion it would generate.

5.1.6 Divergent delivery

Field experimentation in online display advertising presents several challenges to causal inference (Johnson, 2023). In particular, even within A/B tests, advertising platforms' algorithms may optimize campaign delivery over time for predicted user–ad relevance. As a result, different users can be targeted across experimental conditions based on engagement early in the campaign, generating algorithmic selection bias over time (Eckles, Gordon and Johnson, 2018; Braun and Schwartz, 2025). This threat to internal validity—the divergent delivery bias—poses a key concern when the goal is to identify the causal effect of specific post features, rather than the joint effect of algorithmic delivery and post features.

To minimize the risk of divergent delivery and ensure that our estimates reflect the causal effect of the comment section on behavior, rather than the effect of the comment section and platform delivery, we followed and augmented best practices from recent work (Burtch et al., 2025). First, we split budgets evenly across arms, launched all ads simultaneously, and sought to cap exposure at one impression per user. Second, we optimized the campaign for reach rather than engagement. Third, we set a budget large enough to saturate the predefined audience and ran the campaign over multiple weeks to ensure that nearly all users within a defined area were reached. These design features minimize differences in ad delivery across conditions (Braun and Schwartz, 2025). Consequently, any observed differences across conditions can be more confidently interpreted as the causal effect of the comment section on user behavior, rather than as artifacts of algorithmic delivery dynamics.

Since we equalized the number of shares and average reactions across conditions and held the post content constant, the only element that varies across conditions is the comment section. Together with the precautionary measures described above, this design allows us to isolate the effect of the comment section.

5.2 Data

Our data come from two sources. First, we obtain engagement metrics from the Meta Ads Manager dashboard,[20] which reports daily outcomes for each advertisement disaggregated by gender and age group. Second, we collect the full text and timestamps of all user-generated comments directly from the corresponding Facebook posts.

[20] Meta Ads Manager, `https://www.facebook.com/business/tools/ads-manager`.

5.2.1 Outcomes

We analyze a set of engagement outcomes that capture on-platform activity. Specifically, we focus on four metrics: 1) all engagement, defined as any unique user activity related to the ad, including clicks on the ad container, link clicks, profile clicks, reactions, comments, shares, saves, and other observable actions; 2) post expansions, our proxy for attention to the post and its comment section, defined as clicks to expand the ad panel, including clicks to view the comments (this outcome is not directly reported in the Meta Ads Manager interface and is constructed by differencing out other available metrics);[21] 3) interactions, defined as the sum of unique reactions (likes and other emoji responses), comments, and shares (reposting); and 4) unique link clicks, capturing whether a user clicked on the external link, which we interpret as intent to learn more about the campaign.[22,23]

All of these actions can be taken with or without exposure to the comments themselves. Across many platforms, including Facebook, the comment section is not displayed by default. In the feed, users see the post and its aggregate counters, but comment text appears only if they expand the container. Unless otherwise specified, we report the engagement rates computed as the ratio of each outcome to total reach, which is defined as the number of distinct users who were shown the ad at least once.

In addition to these engagement outcomes, we collect the full text and timestamps of all user-generated comments. This allows us to examine not only the volume but also the *stance* of user discourse. We use a large language model (GPT-4) to classify each comment as *supportive* or *non-supportive* of the organization's message, based on its semantic content and sentiment. Analogously, we categorize user reactions according to their valence: "likes," "loves," and "cares" are coded as supportive, while other reaction types (such as "angry," "sad," or "wow") are coded as neutral or not supportive. These additional measures allow us to quantify how pre-existing comment narratives shape the tone and direction of subsequent engagement, thereby linking the stance of visible comments to the ideological composition and sentiment of later user responses.

5.2.2 Sample Characteristics

Appendix Table E2 reports descriptive statistics for the sample used in the main experiment. The final dataset comprises 1,054,015 observations at the user level, the sum of the unique users reached by each of the 72 ads. Treatment assignment is well balanced across arms, with roughly 25 percent of individuals allocated to each of the four conditions (Control, Supportive, Mixed, and

[21] *Post Expansions = Clicks All - Page Engagement. Clicks All* captures all user clicks on an ad, including clicks on the ad container, link clicks, profile clicks, reactions, comments, shares, saves, and other interactions. *Page Engagement* includes all identifiable interactions with the ad or the advertiser's page (link clicks, profile clicks, reactions, comments, shares, saves, and other interactions), except for ad container clicks. The residual therefore isolates clicks on the ad container that expand the post without generating any other observable user activity.

[22] The last three outcomes need not add up exactly to all engagement. For example, a user who both comments and clicks on the link is counted twice when the outcomes are summed but only once in all engagement.

[23] As a robustness check, Section 5.4.3 reports results using landing-page views, which capture downstream off-platform activity. This measure has two limitations. First, landing-page views are inferred via the Facebook Pixel rather than directly recorded by Facebook, which may introduce measurement error. Second, unlike unique link clicks, landing-page views are not unique at the user level.

opposing comments). Engagement outcomes exhibit substantial variation, reflecting the skewed and infrequent nature of user activity on social media platforms. On average, 0.53 percent of reached users engaged with the post in any form, 0.29 percent expanded the ad panel to view the comment section, and 0.24 percent clicked on the external link. Interaction rates—comprising reactions, comments, and shares—averaged 0.02 percent of reach, while 0.18 percent of users visited the organization's landing page.

The reached audience broadly tracks the demographic composition of Meta's U.S. adult audience, although some differences remain. Men account for 52.3 percent of reached users and women for 47.7 percent, compared with 46.8 percent and 53.3 percent, respectively, in Meta's combined U.S. adult (18+) ad audience across Facebook, Instagram, and Messenger. By age, the reached audience is concentrated among users aged 25–44, who represent 57.3 percent of the sample compared with 43.6 percent of Meta's U.S. adult audience, whereas users aged 18–24 and 65+ are relatively underrepresented (8.7 vs. 18.0 percent and 7.5 vs. 12.7 percent, respectively). Taken together, these patterns indicate that the campaign reached a demographically broad segment of U.S. Meta users, but with somewhat greater representation of men and middle-aged users than in the overall adult audience.

5.2.3 Balance Checks

Table 1 reports covariate balance across the four experimental conditions. The sample comprises approximately 1.05 million individuals, evenly distributed across treatment arms, with about 263,000 users per group. We examine three observable individual-level characteristics: gender, middle-aged status (ages 35–64), and senior status (ages 65 and above). Mean values and standard deviations are shown by group, and pairwise differences with the control arm are tested using two-sample t-tests with standard errors clustered at the advertisement level (see Section 5.3 for details).

Across all covariates, differences between treatment and control groups are small in magnitude and statistically insignificant. The p-values from the corresponding tests uniformly exceed conventional significance thresholds, indicating that random assignment produced well-balanced groups across key demographic dimensions. This balance supports the internal validity of the experimental design and suggests that any subsequent differences in engagement outcomes can be attributed to the randomized variation in comment visibility and stance, rather than to pre-existing differences in audience composition.

Appendix Table E3 reports balance checks for ad-level cost and performance metrics across the four experimental conditions. Each observation corresponds to one advertisement, for a total of 72 ads evenly distributed across treatment arms. We compare total spend, cost per mille (CPM), frequency, reach, and spend per user to verify that Meta's delivery algorithm exposed ads in each treatment arm to comparable audience sizes and costs.

Mean values are virtually identical across groups, and none of the pairwise differences relative to the control group are statistically significant. Total spend per ad averages approximately $150, with CPMs around $9.6 and mean reach near 14,600 users. The estimated p-values for all tests are

Table 1: Balance Checks: Individual-level Covariates

Variable	*Group Mean / (SD)*				*t*–test *p*–value		
	(1) Control	(2) Supportive	(3) Mixed	(4) Opposing	(1)–(2)	(1)–(3)	(1)–(4)
Male	0.522 (0.500)	0.522 (0.500)	0.524 (0.499)	0.525 (0.499)	0.982	0.898	0.877
Middle-Aged (35-64)	0.538 (0.499)	0.536 (0.499)	0.537 (0.499)	0.535 (0.499)	0.867	0.944	0.812
Senior (65+)	0.074 (0.262)	0.076 (0.266)	0.075 (0.263)	0.074 (0.262)	0.736	0.937	0.973
Observations	263,706	262,246	263,197	264,866			

Notes: Each observation is a user. Standard errors in the t-tests are clustered at the ad level (72 ads).

well above conventional significance thresholds, confirming that the experimental conditions were implemented under comparable delivery and budget parameters. These results indicate that Meta's optimization algorithm did not differentially allocate impressions or spending across treatment arms, reinforcing the internal validity of our causal design.

The frequency and reach metrics further support the correct implementation of the experimental design. Our objective was to saturate audiences such that each individual would be reached at most once. The estimated average frequency of approximately 1.12 shows that repeat exposure was limited though not eliminated. Combined with an upper-bound estimated audience size of about 14,518 users per ad on average, it is consistent with complete audience saturation and balanced delivery across conditions. Appendix Table E4 reports saturation numbers under different scenarios. Under the most conservative estimate, we reached over 95 percent of the potential audience in the areas we targeted.

5.3 Empirical Strategy

We observe outcomes for 72 ads, one for each combination of audience stratum $z \in \{1, \ldots, 18\}$ and condition $k \in \{\text{No Comments}, \text{Opposing}, \text{Supportive}, \text{Mixed}\}$. Within each stratum, individuals are randomly assigned to one of the four ads. For each ad, we record the number of unique individuals reached and the number who took a given action (e.g., clicking the link or reacting to the post), disaggregated by day, gender, and age group. In the main analysis, we aggregate across days, so each observation covers the full campaign.

To estimate effects at the individual level, we construct a synthetic dataset in which each observation is one individual reached by an ad. The construction is deterministic: each cell defined by ad, gender, and age group is replicated once per individual reached, and $Y = 1$ is assigned to exactly the number of individuals in the cell who took the action. Let p_{zg}^k denote the observed share of individuals in stratum z, condition k, and gender–age group g who took action Y. We model the individual outcome as Bernoulli, $Y_{izg}^k \sim \text{Ber}(p_{zg}^k)$, independent and identically distributed within each cell. Random assignment within strata justifies this assumption. The synthetic data let us

include individual-level covariates and stratum fixed effects, and report coefficients as changes in the probability that a user takes the action.[24]

We then estimate the following linear probability model by ordinary least squares:

$$Y_{iz} = \alpha + \sum_k \beta_k T_i^k + X_{iz}'\gamma + \delta_z + \varepsilon_{iz}, \tag{1}$$

where Y_{iz} is the outcome of individual i in stratum z in the synthetic dataset; T_i^k indicates assignment to condition $k \in \{\text{Opposing}, \text{Supportive}, \text{Mixed}\}$, with *No Comments* as the omitted category; X_{iz} is a vector of gender and age-group indicators, as well as their interactions with the stratum fixed effects; δ_z denotes stratum fixed effects, which the tables label *ZIP Code Set FEs* because each stratum is a set of ZIP codes; and ε_{iz} is an error term. The coefficients β_k capture the causal effect of assignment to each comment condition relative to the control.

We estimate heterogeneous effects by re-estimating equation (1) separately by area ideology. Ideology is measured at the ZIP code group level, not for individuals. A Red area is one where the Republican vote share is high, not one in which every user is conservative, and vote shares correlate with other characteristics such as population density. We therefore interpret these results as heterogeneity across locations.

We cluster standard errors at the stratum–treatment level, corresponding to the 72 ads. This is the level at which the comment section is delivered. Individuals within a cell see the same post and, in the treatment arms, the same comments. Clustering allows arbitrary dependence within cells, so inference does not rely on the i.i.d. assumption behind the synthetic data analysis. It requires that cells be independent of one another. The design supports this: strata are mutually exclusive sets of ZIP codes, ZIP codes used in the comment generation campaign are excluded, and the comment-hiding pipeline prevents users who see the same post from influencing each other. Section 5.4.3 reports robustness to alternative specifications and inference methods, including estimates computed directly on the 72 ad-level rates instead of relying on the synthetic data, and wild cluster bootstrap p-values.

5.4 Results

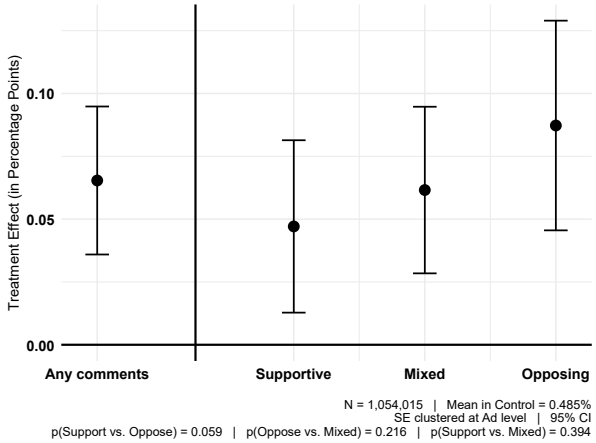
5.4.1 Main Results

We present estimates of the causal impact of the comment section on on-platform user engagement, based on the linear probability model described above, with standard errors clustered at the advertisement level. Figure 4(a) reports the effects on overall engagement, while Figures 4(b–d) present the effects on post expansions, interactions, and link clicks. Effects are reported in percentage points (pp). Full regression results are provided in Appendix Table E5.

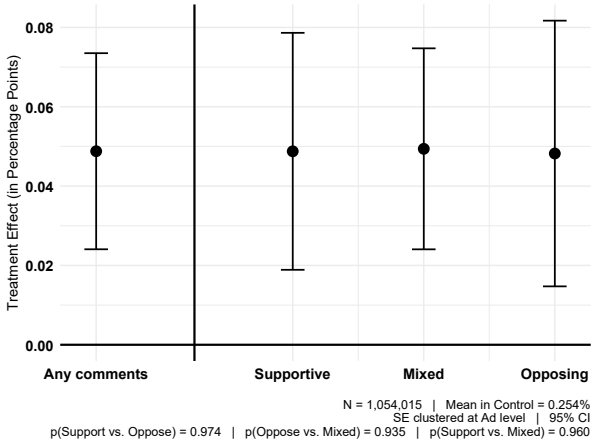
[24]The point estimates are numerically identical to weighted least squares on the cell-level proportions, weighted by cell reach, a standard property of regression on grouped data.

Figure 4: The Impact of the Comment Section on On-platform User Engagement
Outcomes are expressed in % of total reach

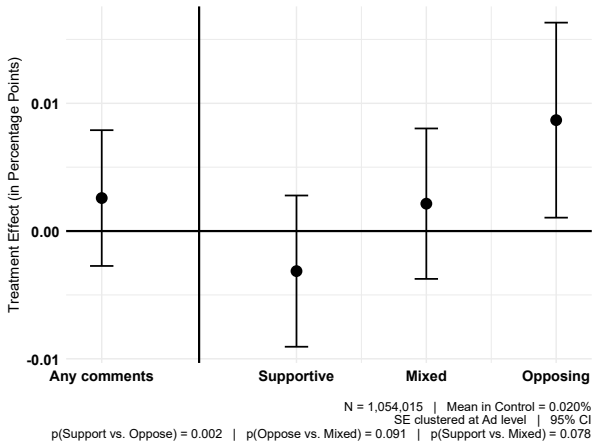
(a) All Engagement



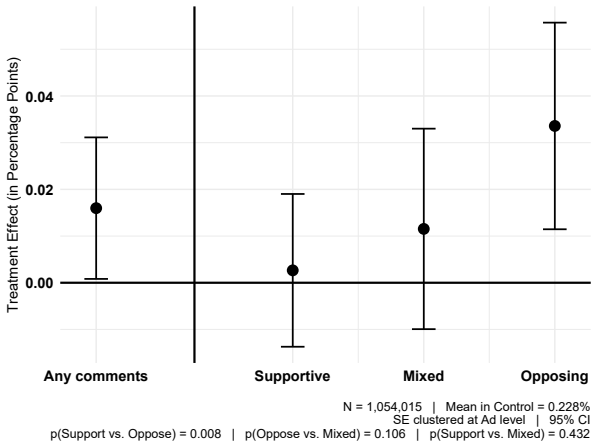
(b) Post Expansions



(c) Interactions (reactions, comments, shares)



(d) Link Clicks



Notes: This figure reports treatment effects of comment visibility and stance on on-platform engagement outcomes, expressed in percentage points of total reach. Panel (a) reports all engagement, panel (b) post expansions, panel (c) interactions (comments, reactions, and shares), and panel (d) unique link clicks. "Any comments" pools the three comment-treatment arms. Estimates are from regression models with the No Comments arm as the omitted category, including gender and age-group controls, ZIP code set fixed effects, and all pairwise interactions among gender, age group, and ZIP code set. Standard errors are clustered at the advertisement level (72 ads). Vertical lines represent 95% confidence intervals. The sample includes 1,054,015 reached users.

All Engagement Figure 4(a) shows that displaying any comments increases total engagement by 0.065 pp ($p < 0.01$), corresponding to a 13.5 percent increase relative to the control mean of 0.485 percent. By stance, opposing comments generate the largest increase (0.087 pp, $p < 0.01$; 18.0 percent), followed by Mixed (0.062 pp, $p < 0.01$; 12.7 percent) and supportive (0.047 pp, $p < 0.01$; 9.7 percent). The difference between opposing and supportive comments is statistically significant at the 10 percent level ($p = 0.059$) and sizable, as the effect of opposing comments is nearly 2 times as large. The results indicate that not only does the presence of comments matter for subsequent engagement, but also their stance. In particular, critical remarks draw substantially more overall

engagement than supportive ones, suggesting that negative commentary tends to amplify overall user activity around the content. We next examine which specific actions drive this pattern.

Post Expansions Figure 4(b) shows that any comments increase the probability that users expand the ad panel to view the comment section by 0.049 pp ($p < 0.01$), corresponding to a 19.2 percent increase over the baseline mean of 0.254 percent. Interpreting post expansions as a proxy for attention, these findings indicate that the presence of comments per se increases attention, independent of their ideological orientation. All three stance conditions produce virtually identical effects—Opposing (0.048 pp, $p < 0.01$), Mixed (0.049 pp, $p < 0.01$), and supportive (0.049 pp, $p < 0.01$)—and none of the pairwise differences are significant ($p > 0.90$). This pattern matches Prediction 1: stance is visible only after the post is expanded, so the decision to open the comment section should not differ across stances. As such, this evidence provides an additional sanity check on the successful randomization for our study, suggesting that differential ad delivery across experimental conditions is unlikely to explain our results.

Interactions Figure 4(c) indicates that the pooled "any comments" effect on reactions, comments, and shares is small and statistically insignificant (0.003 pp). However, stance-specific estimates reveal that opposing comments substantially increase interactions by 0.009 pp ($p < 0.05$), a 43 percent rise relative to the control mean of 0.020 percent. For mixed comments, the coefficient is positive but statistically insignificant (0.002 pp), whereas for supportive comments it is negative and insignificant (-0.003 pp). Differences between opposing and supportive conditions are highly significant ($p < 0.01$), whereas differences involving the Mixed condition are smaller and only marginally significant ($p < 0.10$). This pattern suggests that antagonistic comments elicit greater visible participation from subsequent users.

Link Clicks Figure 4(d) shows that the presence of any comments raises the probability of clicking on the external link by 0.016 pp ($p < 0.05$), representing a 7 percent increase over the baseline mean of 0.228 percent. Opposing comments again produce the largest effect (0.034 pp, $p < 0.01$; 14.8 percent), while the coefficients for Mixed (0.012 pp) and supportive (0.003 pp) comments are smaller and not significant. The difference between opposing and supportive arms is statistically significant ($p < 0.01$), whereas comparisons involving the Mixed condition are not. These results indicate that opposing comments are particularly effective in increasing click-through activity, a key metric for organizations seeking to drive website traffic.

Valence of Interactions The additional engagement generated by opposing comments appears to be predominantly supportive in tone. Comparing effects on all interactions with effects on *supportive interactions*, defined as positive reactions (*likes*, *loves*, *cares*), comments expressing agreement with the organization's message, and shares, opposing comments raise all interactions by 0.0082 pp and supportive interactions by 0.0075 pp (both $p < 0.05$), as shown in Appendix

Figure E2 and Appendix Table E7.[25] Mixed and supportive comment sections have no statistically significant effects, and effects on non-supportive or ambiguous interactions are small and imprecisely estimated. One possibility is that visible disagreement increases the likelihood that users who support the original message choose to express their agreement. At the same time, the initial presence of negative comments may reduce the propensity of other users holding similar views to post additional negative responses. Either way, these patterns suggest that opposing comments alter the composition of subsequent engagement.

Discussion Across all outcomes, the presence of a comment section modestly increases on-platform user engagement. However, much of this overall effect reflects greater attention to the post itself, as captured by post expansions, and does not vary by the stance of existing comments. In contrast, comment stance shapes the composition of subsequent engagement. These patterns are consistent with Prediction 1: the presence of comments increases post expansion regardless of stance, and stance drives subsequent engagement. Opposing comments consistently generate higher rates of interactions and link clicks relative to the No Comments condition, whereas supportive comments do not significantly outperform the control. Differences between the opposing and supportive conditions are substantial, while differences across the remaining conditions are generally small. These engagement gains also lower unit advertising costs: with respect to the control, spend per interaction falls by 50 percent under opposing comments and spend per link click by 15 percent (Appendix Table E6). Taken together, these findings suggest that opposing discourse amplifies participation and interest more than supportive commentary, pointing to a potential trade-off between engagement amplification and polarization in online comment sections. This trade-off echoes findings on other platform features: Germano, Gómez and Sobbrio (2026), for instance, document a similar tension in the context of engagement-maximizing ranking algorithms. In Section 6, we further examine how these engagement patterns affect attitudes and off-platform behavior in a survey experiment.

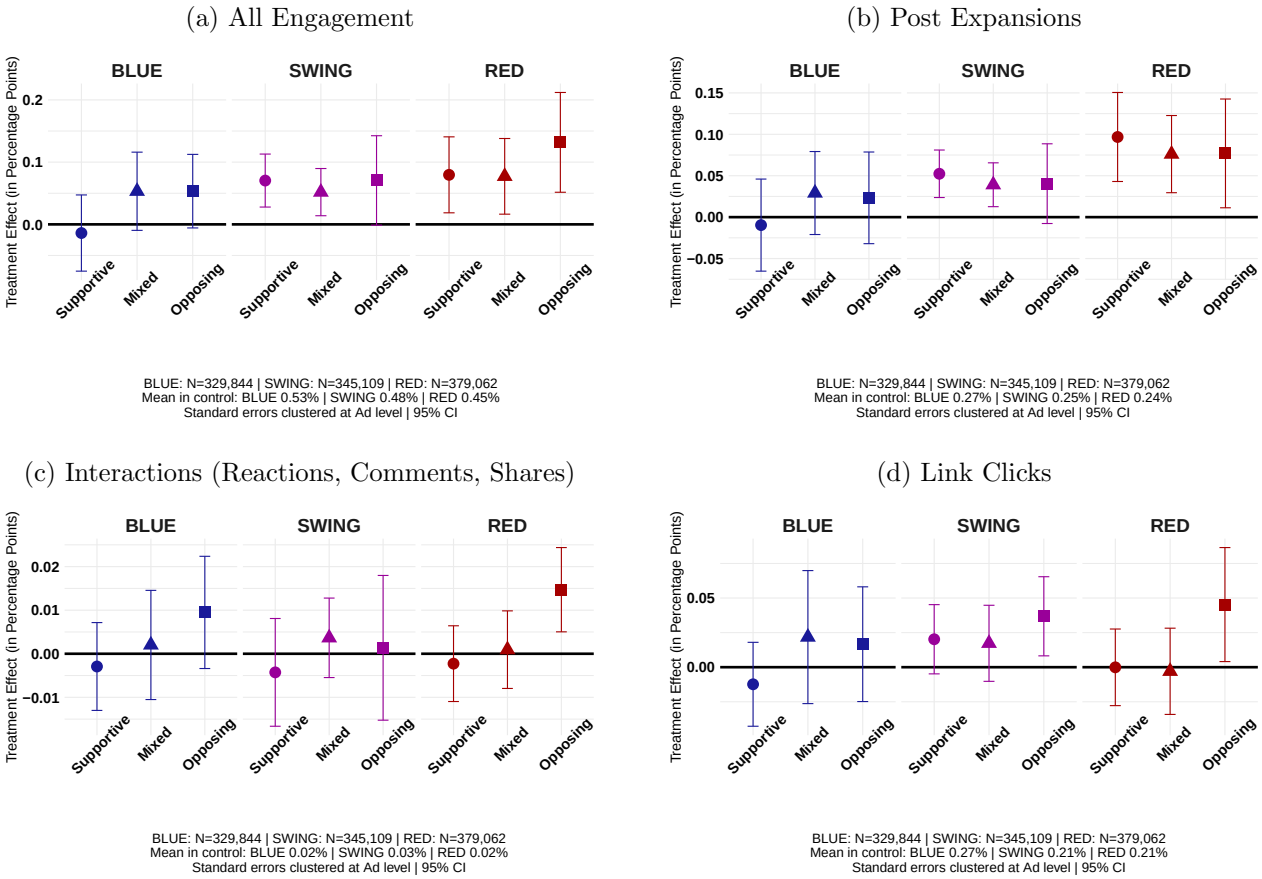
5.4.2 Heterogeneous Effects by Area Ideology

We next examine whether these effects vary with the prevailing political ideology of the area, distinguishing between *Blue* (mostly liberal), *Swing* (mixed), and *Red* (mostly conservative) ZIP codes. Figure 5 and Appendix Tables E8–E10 report the estimates.

The effect of the presence of a comment section rises with the conservativeness of the area. Any comments raise total engagement by 0.031 pp in Blue areas, 0.064 pp in Swing areas ($p < 0.01$), and 0.096 pp in Red areas ($p < 0.01$), or 6, 13, and 21 percent of the respective control means. The estimate for Blue areas is statistically insignificant. In Swing and Red areas, most of this increase consists of post expansions, our proxy for attention, which respond similarly to all three stances.

[25] These estimates are from a specification using data retrieved directly from the posts, without user-level covariates, and with heteroskedasticity-robust standard errors. Results for all interactions are therefore very similar but not identical to those reported above. Minor discrepancies may arise if users deleted their comments or removed reactions after the initial data extraction.

Figure 5: Heterogeneous Treatment Effects by Area Ideology
Outcomes are expressed in % of total reach



Notes: This figure reports heterogeneous treatment effects across areas with different ideological compositions. Panels show treatment effects on all engagement, post expansions, interactions, and unique link clicks, each expressed in percentage points of total reach. BLUE, SWING, and RED areas are defined using Republican vote share cutoffs. Estimates are from regression models estimated separately by area type, with the No Comments arm as the omitted category. Standard errors are clustered at the advertisement level, and vertical lines represent 95% confidence intervals.

Baseline engagement runs in the opposite direction: in the control condition, overall engagement averages 0.53 percent in Blue areas, 0.48 percent in Swing areas, and 0.45 percent in Red areas, consistent with greater alignment between the campaign's message and prevailing ideology in more progressive areas.[26] The comment section thus draws attention in areas where the post itself draws less attention. By making visible that the post has generated discussion, it may increase the salience of content that users in conservative areas would otherwise overlook.

Differences across comment stances are most pronounced in Red areas. There, opposing comments raise total engagement by 0.13 pp ($p < 0.01$), interactions by 0.015 pp ($p < 0.01$), about 2 times the control rate, and link clicks by 0.045 pp ($p < 0.05$), whereas mixed and supportive comments have no detectable effect on interactions or link clicks despite raising post expansions

[26]This pattern is consistent with prior literature (e.g., Song 2024 in the racial justice context), which finds that individuals are more likely to engage with social media content that aligns with their pre-existing attitudes.

as much as opposing comments do. The differences between the Opposing condition and the other conditions are statistically significant for interactions ($p < 0.01$) and link clicks ($p < 0.05$). In Swing areas, all three stances raise total engagement by 0.05–0.07 pp and opposing comments raise link clicks by 0.037 pp ($p < 0.05$), but the stances do not differ significantly from one another. In Blue areas, the estimates are small and imprecise, and only the difference in interactions between opposing and supportive comments is statistically significant ($p < 0.05$).

These patterns suggest that the comment counter raises attention, most where baseline attention to the post is lowest, and once the section is opened, the opposing section is the most salient. Identity congruence may reinforce the latter effect in conservative areas, where an opposing comment both contrasts with the post and aligns with the viewer's own stance, as discussed in Appendix A.3. Because we observe ad-level aggregates, however, we cannot link post expansions to subsequent actions at the user level, and so cannot tell whether the larger stance effects in Red areas arise because more users open the comment section there or because those who open it respond differently to what they see.

5.4.3 Robustness and Placebo Tests

We conduct several robustness checks to assess the sensitivity of our results. First, we re-estimate all main models using alternative combinations of control variables and fixed effects (Appendix Tables E11 and E12), and examine robustness to alternative estimators by employing a logit specification (Appendix Table E13). Second, we compute engagement rates directly at the ad level, without the synthetic data construction, and report the corresponding estimates together with heteroskedasticity-robust t-tests of equality across experimental conditions (Appendix Table E14). The estimates remain robust across all these exercises.

We also check robustness to the inference method, reporting wild cluster bootstrap p-values for all estimates (Cameron, Gelbach and Miller, 2008). Under the wild cluster bootstrap, the effect of any comments on overall engagement and the differences between opposing and supportive comments in interactions and link clicks remain significant at the 5 percent level. Other estimates become less precise, particularly within areas, where only 24 clusters are available (Appendix Tables E5 and E8–E10). Clustering at the ZIP-code-set level leaves our main conclusions unchanged (Appendix Table E15).

We conduct three additional exercises. We first perform a ZIP code exclusion sensitivity analysis, dropping one of the 18 ZIP-code sets at a time and re-estimating the main specifications; the resulting coefficients remain stable, indicating that the findings are not driven by particular geographic areas (Appendix Figures E3-E6). We next repeat this exercise at the level of the stimulus, dropping one stimulus (i.e., 6 ads and more than 80,000 people) at a time, so that each estimate excludes one realization of the comment section. The coefficients are again stable, indicating that the results are not driven by a particular stimulus or comment section (Appendix Figures E7–E10). Moreover, we implement randomization-inference (permutation) tests based on repeated placebo reassignments of treatment status (Young, 2019). For all statistically significant effects in the main

analysis, the observed estimates lie consistently in the tails of the corresponding placebo distributions, with one-sided permutation p-values below 0.02, suggesting that these effects are unlikely to be generated by chance under the original randomization scheme (Appendix Figures E11–E14).

Furthermore, we test the impact of comments on landing-page views, an alternative measure of downstream behavior beyond link clicks. Appendix Table E16 suggests that the presence of any comments increases off-platform engagement by 0.016 percentage points ($p < 0.10$), corresponding to a 9.3 percent increase relative to the control mean of 0.171 percent. Opposing comments yield the largest increase (0.019 percentage points, $p < 0.10$; 11.2 percent), followed by mixed (0.017 percentage points) and supportive comments (0.011 percentage points), both of which are not statistically significant. However, we interpret these findings with caution since landing-page views are inferred via pixel-based tracking and are prone to measurement error.

One potential concern is that the higher engagement driven by opposing comments reflects not their stance per se, but other textual features correlated with stance. In particular, toxicity has been shown to drive up engagement (Beknazar-Yuzbashev et al., 2025). To address this, we classify comments shown in each experimental condition using the Google Perspective API and find that average toxicity levels across the three treatment conditions are similar at around 0.37, so the comments in Opposing are no more toxic than those in other conditions. As a further robustness check, we exclude posts containing comments that exceed the Perspective API's recommended toxicity threshold of 0.7.[27] Appendix Table E17 shows that our results remain robust to this restriction, confirming that the engagement effect of opposing comments is not merely an artifact of toxic content. More broadly, toxicity and stance are conceptually distinct: toxicity is a fixed characteristic of the content itself, whereas the effect of stance is relative to the post. Moreover, identity congruence between the comment and the viewer could play a role. The heterogeneity we document is therefore unlikely to be driven by toxicity or other textual features whose effects do not vary systematically with location ideology.

6 The Impact of Comments on Attitudes and Off-Platform Outcomes

The field experiment provides causal evidence that comment sections influence engagement on the platform. However, it does not allow us to elicit users' beliefs or attitudes. Nor can we directly test the psychological mechanisms, such as emotional reactions driven by ideological congruence between users and content.

To address these limitations, we conduct a complementary survey experiment. This design preserves key features of the organic online environment, including real social media posts, authentic comments collected from Facebook users, and behavioral incentives, while allowing us to measure beliefs, attitudes, and incentivized behaviors at the individual level. By embedding experimentally manipulated comment sections inside a controlled survey setting, we isolate how the presence and stance of comments affect attention, perceptions, attitudes, and incentivized donations.

[27] Perspective API developer documentation, "About the API: Score," `https://developers.perspectiveapi.com/s/about-the-api-score`.

Comments can affect attitudes and behavior through several channels. Our framework in Section 2 predicts that, because comments shift attention away from the post, opposing comments move the audience's attitudes toward the comments and away from the post, while supportive comments, which restate the post, have little effect (Prediction 2). The framework is reduced form, and attitudes and behavior could respond through several other channels. Comments could persuade, if users update on the arguments or information they carry, or backfire, if opposing views trigger reactance (Bail et al., 2018). They could shift perceived norms, if users read the comment section as evidence of what others believe (Bursztyn, González and Yanagizawa-Drott, 2020), or signal the post's credibility (Muchnik, Aral and Taylor, 2013). They may also have little effect, since political and social attitudes are difficult to change (Haaland and Roth, 2023). Whether comments affect attitudes and behavior in practice is therefore an empirical question.

6.1 Design

6.1.1 Recruitment and Sample

We recruited approximately 5,000 participants from Prolific. Eligibility criteria required participants to be U.S. residents aged 18–64, to have completed at least 20 previous Prolific submissions, and to maintain an approval rate of at least 95 percent. Pilot participants were excluded. To enable the study of heterogeneous effects, recruitment was stratified to achieve approximate balance across self-identified political ideology (conservative, moderate, liberal) and gender.[28] From an initial sample of 5,077 eligible respondents, we excluded those who failed the attention check, leaving 3,896 participants. Our final sample consists of the 3,868 respondents who completed the survey.

Appendix Table F1 compares the composition of the analysis sample with the U.S. adult population. Our sample is gender-balanced but more educated, and includes more Republicans and Democrats and fewer independents than the U.S. average. As shown in Appendix Table F2, almost everyone who was randomized completed the survey (0.997 in the control arm, 0.992 under supportive comments, and 0.990 under opposing comments), and completion rates are balanced across experimental arms ($p = 0.134$).[29] The arms differ somewhat in size (1,189, 1,287, and 1,392 respondents in the control, supportive, and opposing arms). These differences are present at assignment, and baseline characteristics are balanced across arms in the analysis sample as shown in Appendix Table F3.

6.1.2 Survey Structure

The survey proceeded in four stages. Participants first completed baseline measures capturing political ideology, racial attitudes, beliefs about others' views, social media usage, and demographics.

[28] Participants were informed that the survey would take approximately 10 minutes and that they could earn a bonus of up to $1 based on performance in incentivized belief elicitation, in addition to being entered into a $100 lottery.

[29] Including respondents who failed the attention check does not qualitatively change the results.

Next, the visibility and stance of the comment section were randomized at the participant level. Participants were then shown three social media posts promoting racial justice, mirroring those used in our Facebook campaigns, with the comment section displayed according to their assigned condition. Finally, participants completed post-exposure measures capturing attention, beliefs, attitudes, behavior, emotional responses, and open-ended questions. To minimize experimenter demand effects and preserve ecological validity, participants were instructed to react to each post as if they encountered it naturally on their Facebook feed.

6.1.3 Treatment Conditions

Participants were randomly assigned with equal probability to one of three conditions. In the *No Comments* control condition, posts were displayed without any visible comment section. In the *Supportive Comments* condition, each post displayed two supportive comments. In the *Opposing Comments* condition, each post displayed two opposing comments.[30]

The posts themselves were identical across conditions. The comments were real comments collected from Facebook users during the comment generation campaign described in Section 4. To protect user privacy, names and profile images were replaced with neutral placeholders. In the comment conditions, we displayed two comments of the same stance, one using a male name and one using a female name, in order to avoid confounding stance with perceived commenter gender. Within a treatment condition, the stance of comments was consistent across all posts shown to a participant. Thus, the only dimension varying across individuals was the presence and stance of the comment section. Appendix Figure F1 provides an example. Each participant viewed three posts in random order, one each on environmental justice, criminal justice and police reform, and education, all under the same condition.

6.2 Outcomes

We measure attention, beliefs, attitudes, behavior, and emotional responses.[31] When multiple measures capture a common construct, we aggregate them into standardized indices using inverse-covariance weighting following Anderson (2008). All component variables are re-oriented so that higher values reflect greater alignment with the organization or post. Results for pre-specified secondary outcomes are reported in Appendix F.3 for completeness.

Time Spent We measure the total time each participant spent viewing the three posts, using both time (in seconds) and $\log(1 + \text{time})$ to reduce the influence of outliers.

Beliefs about Others and Commenters We measure beliefs about others using an incentivized question in which participants estimated the share of participants in the study who agreed with a

[30] As pre-specified in the Pre-Analysis Plan, we omitted the *Mixed* condition from the survey experiment to increase statistical power.

[31] For survey instruments, see Appendix H.

conservative statement on racial inequality. Responses were incentivized using a quadratic scoring rule, with potential earnings of up to $1 based on accuracy relative to the share observed in the study. This measure captures whether comments shift perceptions of others' views, which we interpret as a proxy for perceived social norms.

Attitudes toward Racial Issues and the Organization We construct a Post-Exposure Attitudes Index aggregating measures of attitudes toward the organization, perceived importance of the racial justice issues covered in the posts, willingness to discuss political issues separately with progressives and with conservatives, and attitudes toward Black Americans. All components are re-oriented so that higher values reflect greater alignment with the organization and are combined using inverse-covariance weighting.

Behavioral Outcomes We measure two costly behavioral outcomes. Participants were given the opportunity to subscribe to the organization's mailing list by voluntarily providing their email address. Separately, they were enrolled in a lottery and asked whether they would donate any winnings to Color of Change, and if so, how much. If selected as winners, the chosen amount was deducted from their prize and transferred directly to the organization. We analyze both the extensive margin, defined as an indicator for making a positive donation, and the intensive margin, defined as the committed donation amount.

Curiosity, Reflection, Anger, and Annoyance As a measure of curiosity and reflection, we construct an index combining participants' reported interest in seeing additional comments and their assessments of how thought-provoking they found the post and the comments. Because perception of comments applies only to participants in the treatment conditions, we construct this index for these two groups and compare them. For these participants, we also construct an index combining Likert-scale measures of anger and annoyance triggered by the comments.

6.3 Empirical Strategy

Let Y_i denote an outcome for participant i. We estimate:

$$Y_i = \alpha + \beta_1 \mathit{Supportive}_i + \beta_2 \mathit{Opposing}_i + \tau' X_i + \varepsilon_i, \tag{2}$$

where Y_i is the outcome for respondent i, α is a constant, and $\mathit{Supportive}_i$ and $\mathit{Opposing}_i$ are indicators for assignment to the respective comment conditions, and the omitted category is No Comments. The vector X_i contains pre-specified covariates: baseline racial attitudes, political ideology, beliefs about others' views, and demographics. Standard errors are robust to heteroskedasticity.

The coefficients β_1 and β_2 identify the causal effect of exposure to supportive and opposing comments, respectively, relative to no comments. The contrast $\beta_2 - \beta_1$ captures the differential effect of opposing versus supportive comments.

6.3.1 Descriptive Statistics and Balance Checks

Appendix Figure F2 (Panel a) reports the distribution of responses to the question of how often the respondent reads or checks comments on social media. Approximately 86 percent of survey participants report that they sometimes, often, or very often read comments on social media, and more than 50 percent report doing so often or very often. These patterns are even more pronounced among respondents who report higher social media use (Panel b). This evidence confirms that comment sections are a salient feature of online content consumption and motivates our focus on their effects.

Appendix Table F3 reports baseline characteristics by experimental condition and balance tests. Across all observable covariates, the treatment groups are well-balanced relative to the control group. Differences in means are small in magnitude, and the vast majority of pairwise t-tests fail to reject equality at conventional significance levels, with only a few imbalances in ethnicity, race, and beliefs about others' views. Importantly, joint orthogonality F-tests fail to reject equality across all pairwise comparisons, confirming that treatment assignment is orthogonal to observed baseline characteristics. To account for these marginal imbalances and improve precision, as specified in the pre-analysis plan, we include in the main specifications baseline controls for age, gender, education, race, ethnicity, party affiliation, political views, racial attitudes, and beliefs about others' views. The results are qualitatively unchanged across alternative control specifications.

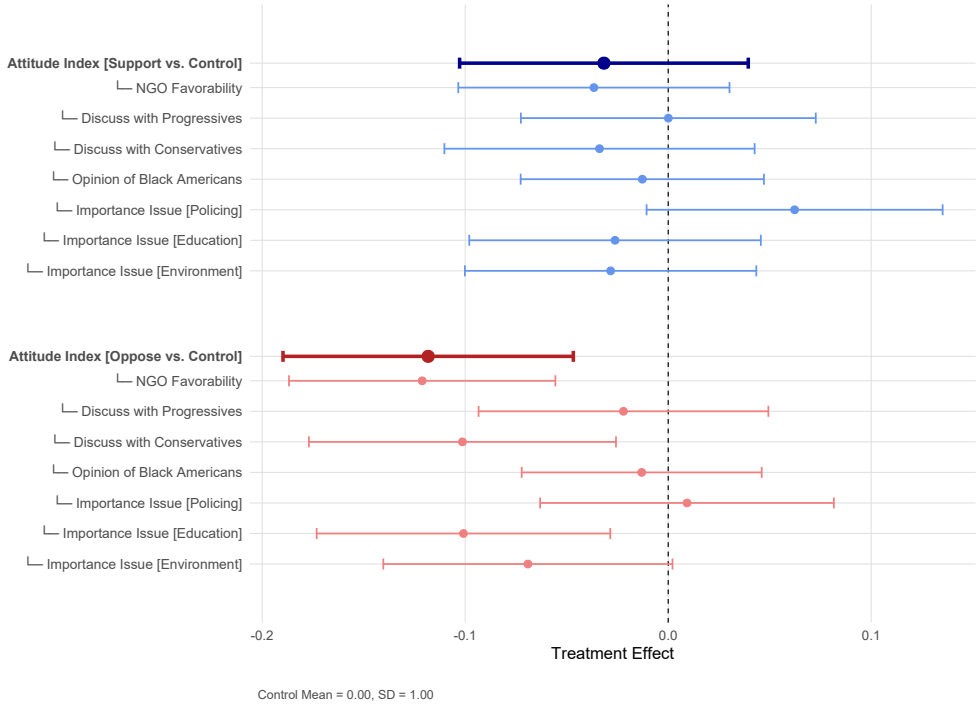
6.4 Results

Time Spent As shown in Appendix Figure F3 and Appendix Table F4, across the three posts shown to respondents, both supportive and opposing comments significantly increase viewing time by approximately 25–27 seconds, a 39–43 percent increase relative to the No Comments condition. These results indicate that people read comments, and the presence of a comment section makes users spend more time on the post regardless of stance. The results are similar when using log viewing time (Appendix Figure F4).

Attitudes Figure 6 shows that opposing comments shift attitudes in a less progressive direction relative to the control condition by approximately 0.12 standard deviations, in line with Prediction 2. The effects are concentrated in measures of NGO favorability and the perceived importance of education equity and environmental justice in the context of racial issues, although the coefficient for the latter is only marginally significant. Opposing comments also significantly reduce the willingness to discuss political issues with conservatives.[32] In contrast, supportive comments have comparatively small and statistically insignificant effects, and the two arms differ significantly ($p = 0.014$). Thus, exposure to opposing comments shifts attitudes away from the organization's

[32]The overall effect remains negative and statistically significant when we exclude the cross-partisan interaction measure from the attitude index. This pre-specified measure can be interpreted as capturing a reduction in affective polarization, though it maps less directly onto progressiveness of attitudes than the other components of the index.

Figure 6: Post-Exposure Racial Attitudes



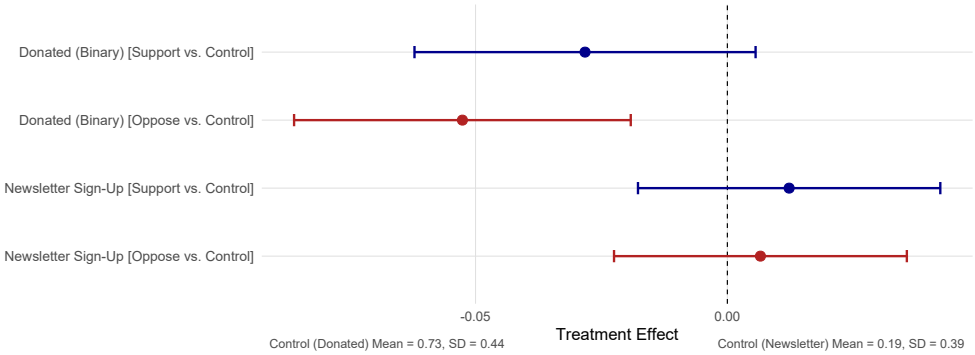
Notes: This figure reports treatment effects of supportive and opposing comments, relative to the control, on post-exposure racial attitudes in the survey experiment. The main outcome is an attitude index, with higher values indicating greater alignment with the organization's position. The figure also reports treatment effects on the index components, including NGO favorability, willingness to discuss with progressives and conservatives, opinions of Black or African Americans, and the perceived importance of policing, education, and environmental issues. Outcomes are standardized to mean zero and standard deviation one in the control group. Horizontal lines represent 95% confidence intervals.

position, with potential implications for support of both the organization and the cause it represents.

Off-platform Behavior As shown in Figure 7, opposing comments significantly reduce the likelihood of making an incentivized donation relative to the control condition by 5.3 percentage points, corresponding to a 7.2 percent decline. They also reduce the amount donated by $2.6, or about 10 percent relative to the control group (Appendix Figure F5). Supportive comments have no statistically significant effect. Neither treatment meaningfully affects newsletter sign-up. Thus, while opposing comments increase attention, they reduce financial support to the organization. Appendix Table F5 reports the corresponding regression estimates for the attitude index and the two behavioral outcomes.

Heterogeneity We further examine heterogeneity in behavioral and attitudinal outcomes by prespecified moderators including baseline racial attitudes, party affiliation, gender, and age (Appendix Figures F6-F9). Overall, we find limited evidence of heterogeneous effects across these dimensions. Estimated treatment effects are broadly similar across groups. In particular, opposing comments reduce willingness to donate for both Democrats and Republicans by 6.4 percent and 12.6 percent,

Figure 7: Donations and Newsletter Sign-Up (Yes/No)



Notes: This figure reports treatment effects of supportive and opposing comments, relative to the control, on off-platform behavioral outcomes in the survey experiment. Outcomes include a binary indicator for newsletter sign-up and a binary indicator for making a positive donation in the incentivized donation task. The control-group mean is 0.73 for donation and 0.19 for newsletter sign-up. Horizontal lines represent 95% confidence intervals.

respectively, relative to the donation rate in the control group.

In contrast, there is substantial heterogeneity in emotional responses. Appendix Figure F10 and Appendix Table F4 show that opposing comments generate higher levels of anger and annoyance, particularly when visible comments conflict with respondents' party affiliation. Relative to supportive comments, opposing comments increase reported anger and annoyance by more than one standard deviation among Democrats, have a smaller positive effect among Independents (about 0.4 standard deviations), and reduce negative emotional responses among Republicans by roughly 0.35 standard deviations. A closely related pattern appears for curiosity and reflection (Appendix Figure F11): relative to supportive comments, opposing comments reduce an index combining curiosity about additional comments and how thought-provoking respondents find the post and its comments by roughly 0.67 standard deviations among Democrats and 0.38 standard deviations among Independents, with no detectable effect among Republicans. These findings indicate that comment stance shapes not only engagement, but also emotional and cognitive reactions, and that identity congruence plays an important role in determining how users respond to visible disagreement.

To provide suggestive evidence on mechanisms behind the attitude and donation effects, we examine beliefs about others by party affiliation. Appendix Figure F12 reports responses to an incentivized belief-elicitation question using a quadratic scoring rule, asking participants what share of survey respondents they believe agreed or strongly agreed with the statement "Black people could be just as well off as white people if they would only try harder". The outcome is reverse-coded, so that higher values indicate more progressive perceived norms. Among Democrats, opposing comments reduce the perceived prevalence of progressive views, consistent with social norms as a potential channel. At the same time, the shift is smaller and insignificant among Republicans and absent among Independents, suggesting that norm updating alone cannot fully account for the effects on attitudes and donations. Other channels, such as persuasion or learning about the

organization's legitimacy, may also play a role.

Discussion The survey evidence reveals a clear tension between engagement and downstream outcomes. Comment sections—especially those featuring opposing remarks—increase engagement but can shift attitudes and reduce financial support, highlighting a trade-off between amplifying engagement and preserving support for the organization and the underlying message.

The experiment on Facebook shows that comment sections on progressive posts are disproportionately populated by conservative voices, and both their presence and stance shape subsequent on-platform behavior. Opposing comments increase clicks and interactions in the field setting, particularly in conservative areas. However, the survey evidence shows that exposure to opposing comments shifts attitudes in a less progressive direction and reduces incentivized donations.

To quantify this trade-off, we conduct a back-of-the-envelope cost-benefit analysis for fundraising campaigns in Appendix G. Under a benchmark calibration that combines the observed increase in click-through rates with the estimated decline in donation propensity, abstracting from changes in delivery efficiency or audience composition, opposing comments can still raise expected donations. However, this result relies on the assumption that the additional users induced to click are no less likely to convert than baseline users. Our evidence suggests that this assumption may be too strong, since opposing comments attract relatively more traffic from less progressive areas, making a deterioration in traffic quality plausible. While this exercise should be interpreted with caution given its simplifying assumptions, it highlights that the net value of tolerating opposing comments depends not only on the increase in traffic they generate, but also on the quality of that traffic.

7 Conclusion

This paper provides causal evidence that comment sections shape both on-platform engagement and downstream attitudes and behavior. In a large-scale Facebook field experiment conducted in collaboration with a leading racial justice organization, we show that the presence of a comment section on progressive posts increases engagement, and that comment stance affects how users respond to content. Opposing comments, in particular, amplify interactions (e.g., comments and reactions) and link clicks relative to supportive comments. Because comment sections are often populated by the vocal few, these findings imply that the visible opinions of a few users can meaningfully shape the on-platform behavior of many others.

Our complementary survey experiment reveals that higher visible engagement does not necessarily imply greater support. While opposing comments increase attention, they also shift attitudes to be less progressive and reduce incentivized donations to the organization. Exposure to counter-attitudinal comments generates anger and annoyance among identity-incongruent users, underscoring that visible disagreement affects not only engagement behavior but also emotional responses and evaluations of the organization. Together, the evidence points to a tension between short-run on-platform engagement gains and potential off-platform longer-run costs.

These results speak to broader concerns about online discourse. Comment sections can amplify divisive narratives in ways that may not reflect the broader audience, potentially complicating the ideal of social media as a deliberative public sphere. At the same time, our findings do not imply that removing comment sections is a straightforward solution: comments increase attention and facilitate participation, and in some contexts may stimulate supportive expression. The challenge lies in how visible discourse is structured and moderated.

For platforms and content producers, moderation involves clear trade-offs. Tolerating contentious or opposing comments can boost engagement and reduce advertising costs, but may undermine brand safety and shift attitudes away from the content producer's objectives. Importantly, platform incentives to maximize engagement may not align with the incentives of firms, nonprofits, or political actors seeking to preserve brand integrity or policy support. Understanding these misalignments is central to current debates over platform governance and content moderation.

A limitation of our study is that our evidence comes from a single organization and a single issue, racial justice. The comment sections in our experiments involve little deliberation and informational content, and those who comment are a small and unrepresentative subset of the audience. This mirrors the concentration of post production on platforms (Grinberg et al., 2019) and is common for news and political issues (Kim and Noh, 2026). We show that under these conditions the vocal few can exert outsized influence, even without any prior connection to the audience they reach. In settings where comments are more deliberative or informative, other mechanisms may play a role and the effects may differ.

Our study points to two promising directions for future research. Methodologically, we introduce a scalable experimental pipeline that manipulates comment visibility and stance using platform-native tools while preserving ecological validity through organic user-generated content. One direction is to extend this framework to other domains, such as commercial products or public health campaigns, to better understand how visible online discourse shapes behavior and to inform the design of comment environments that balance engagement with broader social and organizational goals. A second direction is to move beyond one-time exposure and study how repeated exposure, user interactions, and algorithmic ranking jointly shape engagement and attitudes over time.

References

Acemoglu, Daron, Asuman Ozdaglar, and James Siderius. 2024. "A Model of Online Misinformation." *Review of Economic Studies*, 91(6): 3117–3150.

Anderson, Michael L. 2008. "Multiple Inference and Gender Differences in the Effects of Early Intervention: A Reevaluation of the Abecedarian, Perry Preschool, and Early Training Projects." *Journal of the American Statistical Association*, 103(484): 1481–1495.

Anderson, Monica, Michael Barthel, Andrew Perrin, and Emily A. Vogels. 2020. "#BlackLivesMatter Surges on Twitter after George Floyd's Death." *Pew Research Center*,

https://www.pewresearch.org/short-reads/2020/06/10/blacklivesmatter-surges-on-twitter-after-george-floyds-death/.

Aridor, Guy, Rafael Jiménez-Durán, Ro'ee Levy, and Lena Song. 2024. "The Economics of Social Media." *Journal of Economic Literature*, 62(4): 1422–1474.

Aridor, Guy, Rafael Jiménez-Durán, Ro'ee Levy, and Lena Song. 2026. "Experiments on Social Media." In *Handbook of Experimental Methods in the Social Sciences*. 549–580. Edward Elgar Publishing.

Bail, Christopher A. 2021. *Breaking the Social Media Prism: How to Make Our Platforms Less Polarizing.* Princeton, NJ:Princeton University Press.

Bail, Christopher A, Lisa P Argyle, Taylor W Brown, John P Bumpus, Haohan Chen, MB Fallin Hunzaker, Jaemin Lee, Marcus Mann, Friedolin Merhout, and Alexander Volfovsky. 2018. "Exposure to Opposing Views on Social Media Can Increase Political Polarization." *Proceedings of the National Academy of Sciences*, 115(37): 9216–9221.

Bakshy, Eytan, Dean Eckles, Rong Yan, and Itamar Rosenn. 2012. "Social Influence in Social Advertising: Evidence from Field Experiments." *Proceedings of the 13th ACM Conference on Electronic Commerce (EC)*, 146–161.

Beknazar-Yuzbashev, George, Rafael Jiménez-Durán, and Mateusz Stalinski. 2024. "A Model of Harmful yet Engaging Content on Social Media." *AEA Papers and Proceedings*, 114: 678–683.

Beknazar-Yuzbashev, George, Rafael Jiménez-Durán, Jesse McCrosky, and Mateusz Stalinski. 2025. "Toxic Content and User Engagement on Social Media: Evidence from a Field Experiment." Social Science Research Network Working Paper 5130929.

Berry, George, and Sean J. Taylor. 2017. "Discussion Quality Diffuses in the Digital Public Square." *Proceedings of the 26th International Conference on World Wide Web (WWW)*, 1371–1380.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2012. "Salience Theory of Choice under Risk." *Quarterly Journal of Economics*, 127(3): 1243–1285.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2013. "Salience and Consumer Choice." *Journal of Political Economy*, 121(5): 803–843.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2018. "Diagnostic Expectations and Credit Cycles." *The Journal of Finance*, 73(1): 199–227.

Bordalo, Pedro, Nicola Gennaioli, and Andrei Shleifer. 2022. "Salience." *Annual Review of Economics*, 14: 521–544.

Braun, Michael, and Eric M Schwartz. 2025. "Where A/B Testing Goes Wrong: How Divergent Delivery Affects What Online Experiments Cannot (and Can) Tell You About How Customers Respond to Advertising." *Journal of Marketing*, 89(2): 71–95.

Bursztyn, Leonardo, Alessandra L González, and David Yanagizawa-Drott. 2020. "Misperceived Social Norms: Women Working Outside the Home in Saudi Arabia." *American Economic Review*, 110(10): 2997–3029.

Burtch, Gordon, Robert Moakler, Brett R Gordon, Poppy Zhang, and Shawndra Hill. 2025. "Characterizing and Minimizing Divergent Delivery in Meta Advertising Experiments." arXiv Working Paper arXiv:2508.21251. https://arxiv.org/abs/2508.21251.

Cameron, A. Colin, Jonah B. Gelbach, and Douglas L. Miller. 2008. "Bootstrap-Based Improvements for Inference with Clustered Errors." *The Review of Economics and Statistics*, 90(3): 414–427.

Donati, Dante, and Lena Song. 2024. "The Impact of Comments on Social Media Campaigns." *AEA RCT Registry. July 22*, https://doi.org/10.1257/rct.13812-2.2.

Donati, Dante, and Lena Song. 2026. "The Impact of Social Media Comments on Beliefs, Attitudes, and Behavior." *AEA RCT Registry. February 06*, https://doi.org/10.1257/rct.17850-1.1.

Donati, Dante, and Nandan Rao. 2025. "Adaptive Survey Sampling via Ad Platforms." Social Science Research Network Working Paper 5495148. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5495148.

Donati, Dante, Nandan Rao, Victor Orozco-Olvera, and Ana María Muñoz Boudet. Forthcoming. "Can Facebook Ads Prevent Malaria? Two Field Experiments in India." *Marketing Science*.

Eckles, Dean, Brett R Gordon, and Garrett A Johnson. 2018. "Field Studies of Psychologically Targeted Ads Face Threats to Internal Validity." *Proceedings of the National Academy of Sciences*, 115(23): E5254–E5255.

Gauthier, Germain, Roland Hodler, Philine Widmer, and Ekaterina Zhuravskaya. 2026. "The Political Effects of X's Feed Algorithm." *Nature*, 652(8109): 416–423.

Germano, Fabrizio, Vicenç Gómez, and Francesco Sobbrio. 2026. "Ranking for Engagement: How Social Media Algorithms Fuel Misinformation and Polarization." *Journal of Public Economics*, 255: 105589.

Gordon, Brett R, Florian Zettelmeyer, Neha Bhargava, and Dan Chapsky. 2019. "A Comparison of Approaches to Advertising Measurement: Evidence from Big Field Experiments at Facebook." *Marketing Science*, 38(2): 193–225.

Grinberg, Nir, Kenneth Joseph, Lisa Friedland, Briony Swire-Thompson, and David Lazer. 2019. "Fake News on Twitter During the 2016 U.S. Presidential Election." *Science*, 363(6425): 374–378.

Guess, Andrew M, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, et al. 2023*a*. "How Do Social Media Feed Algorithms Affect Attitudes and Behavior in an Election Campaign?" *Science*, 381(6656): 398–404.

Guess, Andrew M, Neil Malhotra, Jennifer Pan, Pablo Barberá, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, Deen Freelon, Matthew Gentzkow, et al. 2023*b*. "Reshares on Social Media Amplify Political News but Do Not Detectably Affect Beliefs or Opinions." *Science*, 381(6656): 404–408.

Haaland, Ingar, and Christopher Roth. 2023. "Beliefs about Racial Discrimination and Support for Pro-Black Policies." *Review of Economics and Statistics*, 105(1): 40–53.

Johnson, Garrett A. 2023. "Inferno: A Guide to Field Experiments in Online Display Advertising." *Journal of Economics & Management Strategy*, 32(3): 469–490.

Kemp, Simon. 2025. "Digital 2025: The United States of America." DataReportal, Kepios, We Are Social, and Meltwater Report. `https://datareportal.com/reports/digital-2025-united-states-of-america`. Based on data published in Meta's ad planning tools, January–February 2025; accessed April 20, 2026.

Kim, Sangbeom, and Seonhye Noh. 2026. "Disproportionate Voices: Participation Inequality and Hostile Engagement in News Comments." *Proceedings of the International AAAI Conference on Web and Social Media*, 20(1): 1256–1272.

Lee, Dokyun, Kartik Hosanagar, and Harikesh S. Nair. 2018. "Advertising Content and Consumer Engagement on Social Media: Evidence from Facebook." *Management Science*, 64(11): 5105–5131.

Levy, Ro'ee. 2021. "Social Media, News Consumption, and Polarization: Evidence from a Field Experiment." *American Economic Review*, 111(3): 831–870.

Luca, Michael. 2015. "User-generated content and social media." In *Handbook of media Economics*. Vol. 1, 563–592. Elsevier.

Moehring, Alex. Forthcoming. "Personalization, Engagement, and Content Quality on Social Media: An Evaluation of Reddit's News Feed." *Management Science*.

Muchnik, Lev, Sinan Aral, and Sean J Taylor. 2013. "Social Influence Bias: A Randomized Experiment." *Science*, 341(6146): 647–651.

Nyhan, Brendan, Jaime Settle, Emily Thorson, Magdalena Wojcieszak, Pablo Barberá, Annie Y Chen, Hunt Allcott, Taylor Brown, Adriana Crespo-Tenorio, Drew Dimmery, et al. 2023. "Like-Minded Sources on Facebook Are Prevalent but Not Polarizing." *Nature*, 620(7972): 137–144.

Pew Research Center. 2024. "Racial Attitudes and the 2024 Election." *Web report*, `https://www.pewresearch.org/politics/2024/06/06/racial-attitudes-and-the-2024-election/`, accessed December 4, 2025.

Song, Lena. 2024. "Closing the Distance: The Effects of Social Media Content on Support for Racial Justice." Social Science Research Network Working Paper 4832220. `https://papers.ssrn.com/sol3/papers.cfm?abstract_id=4832220`.

Tucker, Catherine E. 2014. "Social Networks, Personalized Advertising, and Privacy Controls." *Journal of Marketing Research*, 51(5): 546–562.

Young, Alwyn. 2019. "Channeling Fisher: Randomization Tests and the Statistical Insignificance of Seemingly Significant Experimental Results." *The Quarterly Journal of Economics*, 134(2): 557–598.

Zhuravskaya, Ekaterina, Maria Petrova, and Ruben Enikolopov. 2020. "Political Effects of the Internet and Social Media." *Annual Review of Economics*, 12: 415–438.

Online Appendix

The Influence of the Vocal Few: Evidence from Social Media Comments

Dante Donati and Lena Song

A Framework Appendix

This appendix states the setup and results of the framework described in Section 2.

A.1 Setup

A post is shown either without comments, $c = \emptyset$, or with a section that is supportive or opposing, $c \in \{S, O\}$. Define $x(c)$ as the attitude the section expresses toward the organization and its cause, measured relative to the post's own position, with

$$x(\emptyset) = x(S) = 0, \qquad x(O) = -1.$$

A supportive section repeats the post's position, while an opposing section expresses a different view and therefore provides the greatest contrast.

As in our empirical setting and on many other platforms, we assume that a comment counter is visible next to a post before the comment section is opened, making the post and the presence of comments more prominent. Denote this prominence by $\phi(c)$, measured relative to a post with no comments and common to S and O, with $\phi(\emptyset) = 0$ and $\phi(S) = \phi(O) = \phi > 0$.

When comments are present, the viewer expects the section to be opposing with probability $\widehat{q} \in (0, 1)$, which depends on her belief about who comments: a viewer who thinks progressives do most of the commenting below a progressive post holds low $\widehat{q}$.

Applying the three channels through which a stimulus captures attention (Bordalo, Gennaioli and Shleifer, 2022), the salience of section c is

$$s(c) = \underbrace{\phi(c)}_{\text{prominence}} + \underbrace{|x(c)|}_{\text{contrast}} + \underbrace{\gamma\big(-\log p_c\big)}_{\text{surprise}}, \qquad \gamma \geq 0, \tag{3}$$

where p_c is the probability the viewer, having seen the counter but not yet the comments, assigns to the section being of stance c. Surprise is strictly decreasing in p_c: the less a section was expected, the more salient it is. A counter showing no comments reveals $c = \emptyset$, so $p_\emptyset = 1$ and $s(\emptyset) = 0$. When the counter shows comments, $p_S = 1 - \widehat{q}$ and $p_O = \widehat{q}$, so

$$\begin{aligned} s_S &= \phi + 0 + \gamma\big[-\log(1-\widehat{q})\big], \\ s_O &= \phi + 1 + \gamma\big[-\log \widehat{q}\big], \end{aligned} \qquad \text{so} \qquad s_O - s_S = 1 + \gamma \log \frac{1-\widehat{q}}{\widehat{q}}. \tag{4}$$

An opposing section always has higher contrast, and it is also more surprising if the viewer expected a supportive discussion, $\widehat{q} < 1/2$.

Let $a(s)$ denote the attention the post and its comment section jointly receive, and $w(s) \in [0, 1]$ the share of that attention devoted to the comment section, both strictly increasing in s.

Engagement. Because stance is not visible until the post is expanded, the viewer decides in two stages. Let R indicate that she expands the post, which reveals the comment section if there is

one, and let $A(c)$ indicate that she takes some further action, such as reacting or clicking a link, when the post is shown with section c. Let G_R and G_A be strictly increasing response functions that map the level of attention into the probability of each action. This captures the assumption that greater attention makes a response more likely by increasing the opportunity to process and act on the stimulus. Then

$$\Pr(R = 1 \mid c) = G_R\big(a(\phi(c))\big), \qquad \Pr\big(A(c) = 1 \mid R\big) = \begin{cases} G_A\big(a(\phi(c))\big) & \text{if } R = 0, \\ G_A\big(a(s(c))\big) & \text{if } R = 1. \end{cases} \tag{5}$$

Before the viewer expands the post, a post with a positive comment counter is more salient than a post without comments through prominence, which raises the probability of expanding the post, $G_R(a(\phi)) > G_R(a(0))$, equally for S and O. A viewer who does not open the section can still act in response to the post, in which case the level of attention is $a(\phi(c))$. For a viewer who opens the section, the salience of its content also enters, so attention is $a(s(c))$ and the probability of subsequent engagement is $G_A(a(s(c)))$.

Attitudes. Let $\mu(c)$ denote the change in the viewer's attitude toward the organization and its cause relative to seeing the post alone. The comment section moves it by shifting weight from the post toward the section's own position, in proportion to the attention each receives, so that salience enters as a decision weight distorted toward what draws the eye (Bordalo, Gennaioli and Shleifer, 2012, 2013):

$$\mu(c) = \big(1 - w(s(c))\big) \cdot 0 + w(s(c))\, x(c) = w\big(s(c)\big)\, x(c). \tag{6}$$

Equation (6) assumes that comments move attitudes only through how far the position they express departs from the post, so supportive comments, which restate the post, leave attitudes unchanged. If supportive comments instead reinforced the post or shifted perceived social norms, they could also move attitudes, as discussed in Section 6.4.

A.2 The salience trade-off

Lemma 1 (The salience trade-off)**.** Suppose that the viewer expects a supportive discussion, $\widehat{q} < 1/2$, so that $s_O > s_S$. Relative to a supportive section, an opposing section

(i) leaves the probability of opening unchanged but, conditional on opening, strictly raises the probability of subsequent engagement, $\Pr\big(A(O) = 1 \mid R = 1\big) = G_A(a(s_O)) > G_A(a(s_S)) = \Pr\big(A(S) = 1 \mid R = 1\big)$, and

(ii) strictly lowers the attitude, $\mu(O) = -w(s_O) < 0 = \mu(S)$, which holds for any $\widehat{q}$ since $x(S) = 0$.

Proof. Write $g(\widehat{q})$ for $s_O - s_S$ in equation (4). If $\gamma = 0$ then $g \equiv 1 > 0$. If $\gamma > 0$ and $\widehat{q} < 1/2$, then $(1 - \widehat{q})/\widehat{q} > 1$, so the logarithm is positive and $g(\widehat{q}) > 1 > 0$. In either case $s_O > s_S$.

For (i), the opening probability is $G_R(a(\phi))$ under both stances because the comment content is not yet visible. If $R = 0$, the probability of a subsequent action is $G_A(a(\phi))$ under both stances.

If $R = 1$, strict monotonicity of a and G_A, together with $s_O > s_S$, imply $G_A(a(s_O)) > G_A(a(s_S))$. Thus, opposing comments raise the unconditional probability of subsequent engagement whenever the section is opened with positive probability. The difference is $G_R(a(\phi))\big[G_A(a(s_O)) - G_A(a(s_S))\big] > 0$.

For (ii), equation (6) gives $\mu(c) = w(s(c))\,x(c)$. Since $x(S) = 0$, $\mu(S) = 0$ for any w. Since $x(O) = -1$, $\mu(O) = -w(s_O) < 0$. □

By (5), opening depends on prominence alone, so expansions rise with the presence of comments but do not differ across stances. Once the comment section is open, greater salience raises the level of attention and hence the probability of a subsequent action. By (6), the same rise in salience shifts the allocation of attention toward a section carrying $x(O) = -1$, pulling the evaluation away from the post, while reweighting a supportive section with $x(S) = 0$ leaves the evaluation unchanged. The stance that generates the most engagement is therefore the one that moves attitudes against the post.

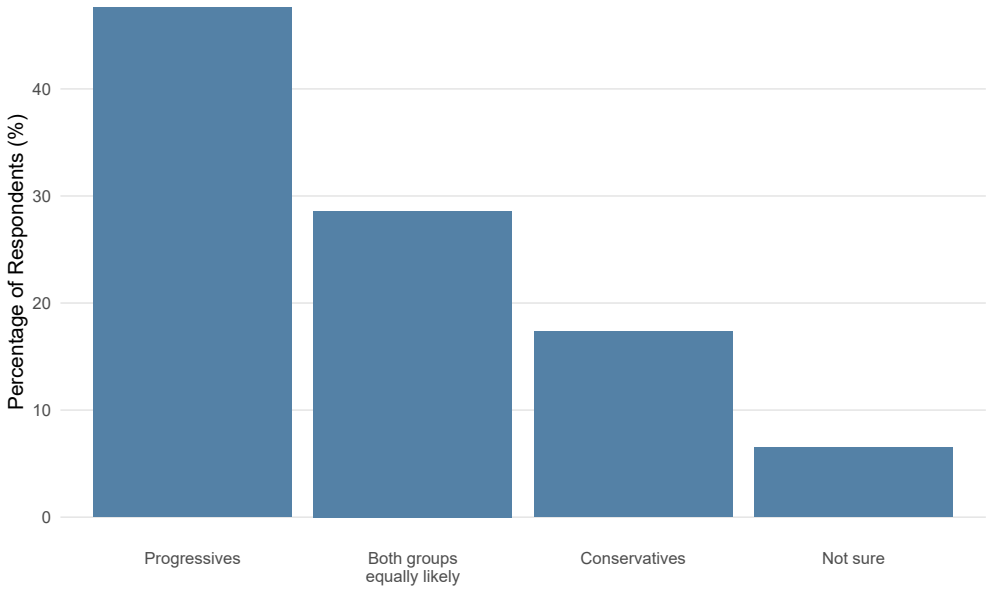
The condition $\widehat{q} < 1/2$ is sufficient but not necessary. If $\gamma = 0$, then $s_O > s_S$ for every $\widehat{q}$. If $\gamma > 0$, then by equation (4), $s_O > s_S$ holds if and only if

$$\widehat{q} < \overline{q}(\gamma) \equiv \frac{1}{1 + e^{-1/\gamma}} > \tfrac{1}{2},$$

so the ordering reverses only once surprise runs against opposing content strongly enough to offset the contrast effect.

Appendix Figure A1 reports a related belief measured in the survey experiment. Among respondents in the control group, 47.6% expect progressives to be the more likely commenters and only 17.3% expect conservatives. Most respondents therefore do not expect conservatives to be more likely to comment, even though users in conservative areas comment at higher rates as shown in Section 4, consistent with them not anticipating an opposing comment section.

Figure A1: Expected Ideology of Commenters in the Control Condition



Notes: This figure reports beliefs about who comments among the 1,189 survey respondents assigned to the control condition, who saw the post without a comment section. Respondents were asked which group they believe is more likely to comment on the post.

A.3 Discussion

Equation (3) should be interpreted as a reduced-form application of salience. In Bordalo, Gennaioli and Shleifer (2012) and Bordalo, Gennaioli and Shleifer (2013), salience is defined relative to payoffs or attributes in a choice set. Here the reference is the post, which is what surrounds the comments on the screen, and the surprise term $\gamma(-\log p_c)$ is an addition motivated by the broader account in Bordalo, Gennaioli and Shleifer (2022) of attention being drawn to what is unexpected.

The framework abstracts from viewer ideology. Salience is defined relative to the post, whereas identity congruence is defined relative to the viewer. A comment opposing a progressive post contrasts with the post for every viewer, but aligns with a conservative viewer's own position and conflicts with a progressive viewer's. The same comment can therefore invite approval from one viewer and anger from another. An extension of the model is to let the payoff of acting depend on identity congruence. This may explain why differences across comment stances are most pronounced in conservative areas: for a conservative viewer, salience and identity congruence both push toward engaging with an opposing section, whereas for a progressive viewer they may move in opposite directions.

B Pre-registration and Pre-analysis Plan

Both Facebook and Prolific experiments were pre-registered in the American Economic Association Registry for randomized controlled trials (under AEARCTR-0013812 and AEARCTR-0017850), be-

fore the corresponding experiment was fielded. We follow the two plans in the experimental designs, audience construction and stratification, randomization, treatment conditions, post and comment selection, estimating equations, covariate sets, and index construction. The field experiment plan also pre-specified heterogeneity analyses by political ideology, racial composition, age, and gender. We report heterogeneity by the political ideology of the area; the analyses by racial composition, age, and gender are available upon request.

Two outcome-level deviations apply to the field experiment. First, we add post expansions, which was not pre-specified, as our proxy for seeing the comment section. Second, we do not report treatment effects on newsletter sign-ups recorded by the organization's website as the campaign generated too few sign-ups to support inference on this outcome, so we instead report unique link clicks. We also compare conditions using the linear probability model in equation (1), together with model-free comparisons of reach-weighted means and heteroskedasticity-robust t-tests (Appendix Table E14), rather than the χ^2 tests of proportions named in the plan. Finally, the realized sample is also smaller than anticipated: the plan targeted 2,500,000 users reached, whereas the campaign reached 1,054,015 across 18 ZIP code sets. Reach per ZIP code set was in line with the plan, so the shortfall reflects the number of sets fielded rather than under-delivery within them.

For the survey experiment, the realized analysis sample of 3,868 respondents is smaller than the 4,800 anticipated in the plan, primarily because fewer conservatives were available on Prolific than the ideology quotas required. Descriptive analyses specified in the plan, including those based on the open-response questions, are available upon request.

C Background Appendix

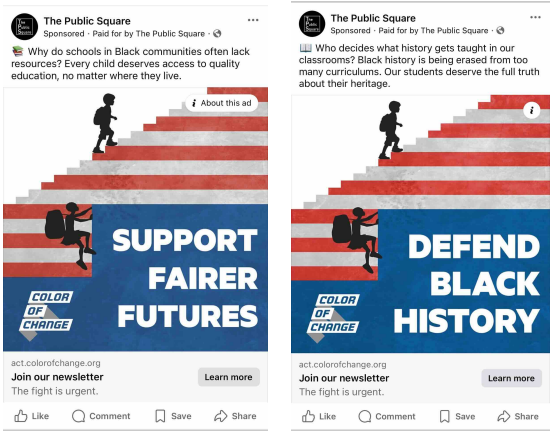
C.1 List of Issues and Description

Here are the racial justice issues and their descriptions presented to content designers to guide the creation of posts for each issue:

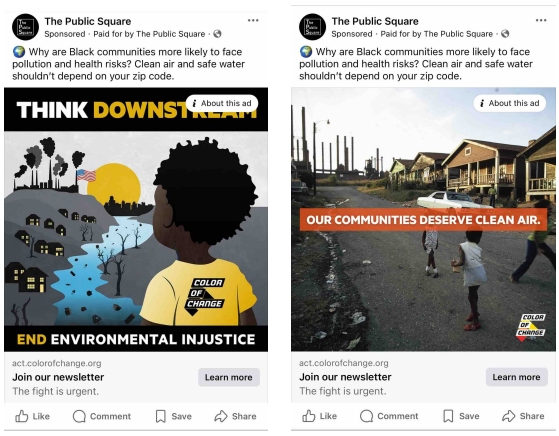
- Voter Suppression: Black communities face deliberate barriers like restricted polling access, strict ID laws, and voter roll purges, aimed at limiting their voting power. Misinformation campaigns also target Black voters to reduce turnout, undermining fair representation. Breaking down these barriers is crucial to ensure Black voices are heard in democratic processes.
- Environmental Justice: Black communities often live near pollution sources like factories and highways, leading to higher rates of health issues such as asthma. These neighborhoods are frequently overlooked in clean-up efforts and lack green spaces. Environmental justice aims to provide Black communities with clean air, safe water, and healthy environments.
- Criminal Justice and Police Reform: Black communities experience disproportionate police violence, profiling, and harsher sentencing. This systemic bias erodes trust in law enforcement and perpetuates disadvantages. Police reform is essential for fair treatment, accountability, and ensuring Black communities feel protected, not targeted, by the justice system.

- Education Reform: Black students often attend underfunded schools with fewer resources, larger classes, and limited access to advanced courses. These disparities create achievement gaps and limit future opportunities. Education reform seeks equitable funding and support to provide Black students with the quality education they deserve.
- Technology Fairness: Black communities face systemic biases in technology, from algorithmic discrimination in hiring and lending to facial recognition tools that disproportionately misidentify Black individuals. These inequities perpetuate existing racial disparities and limit opportunities. Ensuring technology fairness involves designing inclusive systems, addressing bias in algorithms, and creating tools that serve all communities equitably.

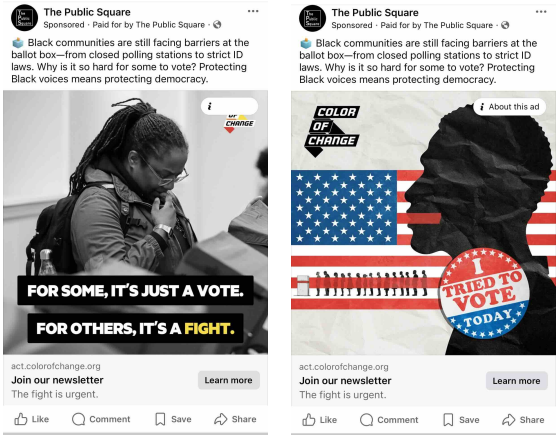
Figure C1: Ad Banners and Headlines



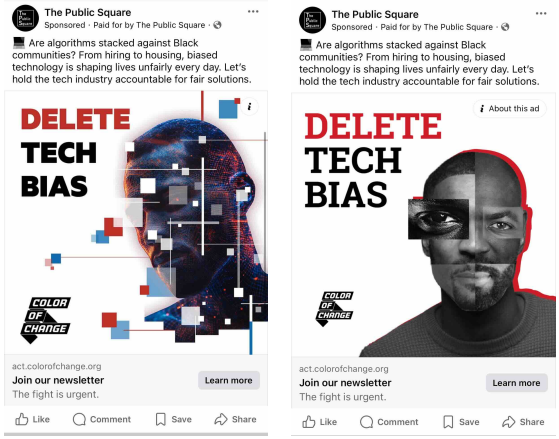
(a) Education



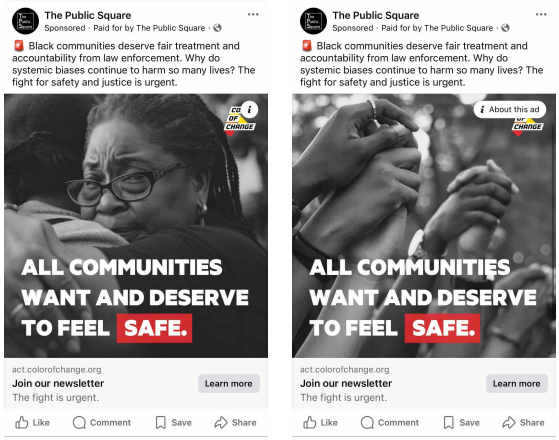
(b) Environment



(c) Voting



(d) Technology



(e) Police

Notes: This figure displays the ad banners and headlines used in the pre-tests.

C.2 Pre-tests

We conducted several pre-tests to systematically select the posts used in the study and to refine the campaign parameters. In Pre-test A, we used Facebook's A/B testing tool across all 10 banners to identify, within each issue, which posts were most likely to generate a high number of clicks.[33] Table C1 summarizes the click-through rates (CTRs)—the ratio of link clicks over reach—for this test. These vary between 0.12% and 0.25%. For each issue, the banners with higher CTR are displayed on the right in Appendix Figure C1.

We conducted two additional tests, Pre-tests B and C, where we capped the frequency at one impression per person. Test B was conducted with a large potential audience (approximately 200,000 users per banner), while Test C targeted a smaller potential audience (approximately 6,000 users per banner). These adjustments were made to simulate a campaign that closely resembles the one planned for our main experiment.

Table C2 presents the aggregate results for Pre-tests B and C.[34] Notably, the CTRs for Pre-tests B and C are significantly lower than those reported for Pre-test A. This discrepancy arises because Pre-test A did not impose a frequency cap, allowing users to see each banner an average of two times and thereby increasing the likelihood of clicks. By contrast, Pre-tests B and C adopt a configuration similar to that used in the main experiment, in which we saturate an audience group by imposing a frequency cap of one impression per user. While this approach may result in lower CTRs, it is essential to mitigate potential divergent delivery bias in Facebook A/B tests caused by the ad platform's algorithm (Burtch et al., 2025), as further discussed in Section 5.1.

Table C1: Results from Pre-test A

Issue	Ad Name	Link Clicks	Reach	CTR (%)
technology	pixels	15	6115	0.245
technology	man	13	6683	0.195
voting	lady	14	5777	0.242
voting	flag	11	5792	0.190
police	lady	13	6146	0.212
police	hands	13	6315	0.206
environment	kid	13	6235	0.209
environment	street	11	6290	0.175
education	future	9	6216	0.145
education	history	6	4884	0.123

Notes: This table reports link clicks, reach, and click-through rates (CTR) for each ad creative tested in Pre-test A, used to select the best-performing banner and headline for each issue prior to the main experiment.

[33] In Pre-test A, we specified the audience (ZIP codes with a high share of progressive populations), budget ($50 per banner), duration (1 week), and optimization goal (reach), without imposing a frequency cap.

[34] The banners on education were excluded from these tests due to their low performance in Pre-test A.

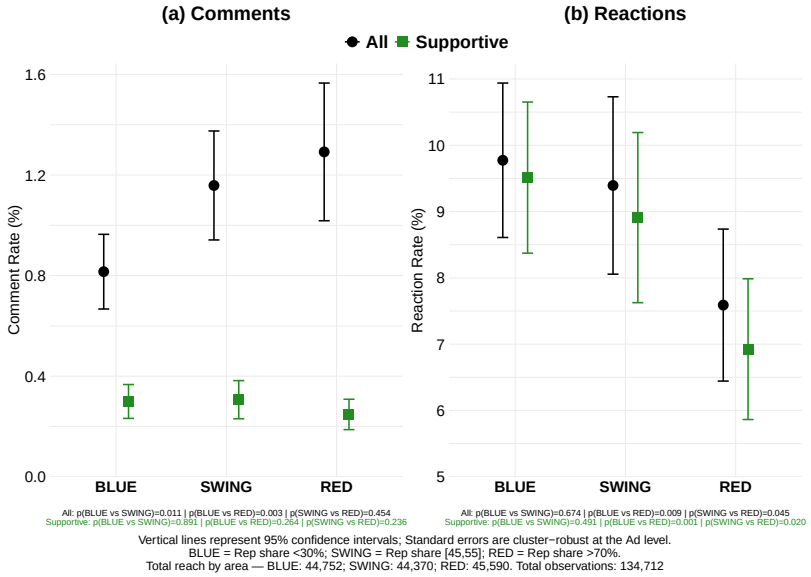
Table C2: Results from Pre-tests B and C

Issue	Ad Name	Link Clicks	Reach	CTR (%)
voting	lady	32	26794	0.119
environment	kid	33	28351	0.116
police	hands	32	27771	0.115
technology	pixels	29	28208	0.103

Notes: This table reports link clicks, reach, and click-through rates (CTR) for each ad creative tested in Pre-tests B and C. Link clicks and reach are pooled across the two tests.

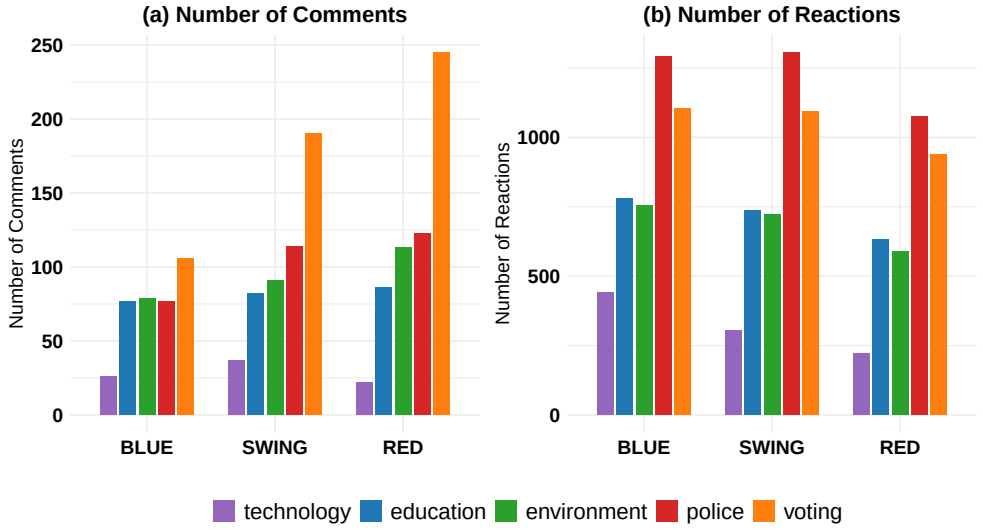
D Descriptive Evidence on Engagement: Additional Results

Figure D1: Comment and Reaction Rates by Valence and Location



Notes: This figure reports comment and reaction rates, expressed as percentages of total reach, by location. Black markers denote all comments/reactions and green markers denote supportive comments/reactions. Supportive reactions include likes, loves, and cares. Vertical lines represent 95% confidence intervals, and standard errors are cluster-robust at the advertisement level. Total reach is 44,752 in BLUE areas, 44,370 in SWING areas, and 45,590 in RED areas.

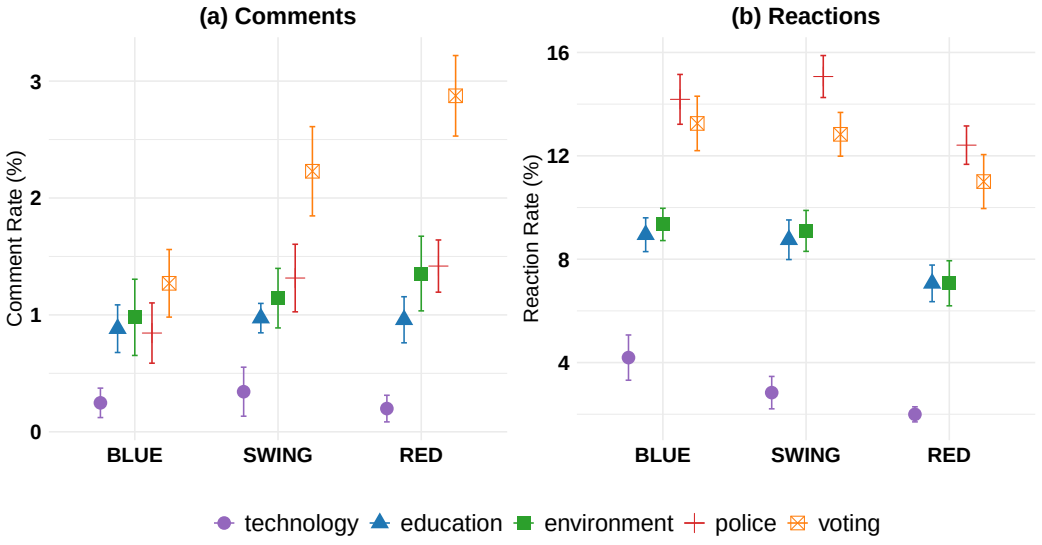
Figure D2: Comment and Reaction Counts by Issue and Location



Bars indicate the total number of comments/reactions.
Observations — technology: 32,350, education: 26,135, environment: 24,375, police: 26,458, voting: 25,394

Notes: This figure reports the total number of comments and reactions generated during the initial Facebook campaign, by issue and location. Colors denote issue areas: technology fairness, education reform, environmental justice, criminal justice and police reform, and voting rights. Areas are grouped as BLUE, SWING, and RED according to Republican vote share.

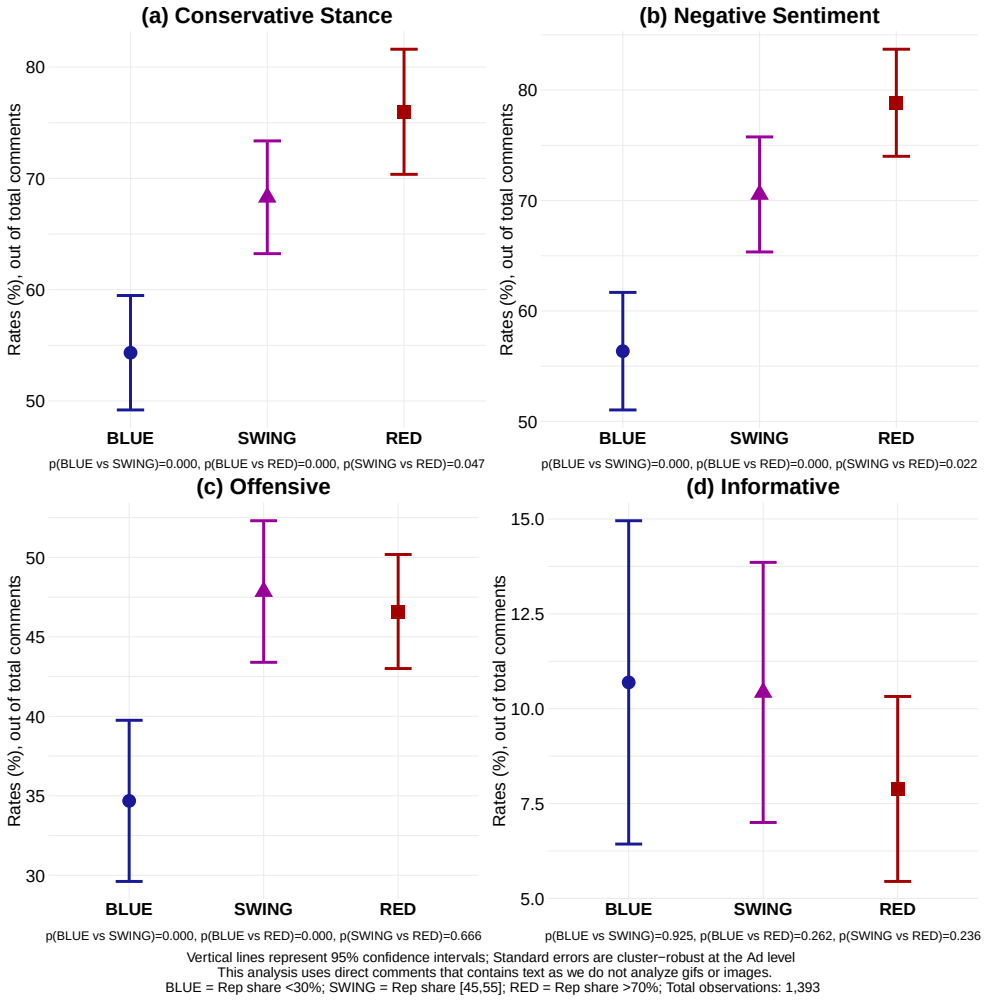
Figure D3: Comment and Reaction Rates by Location and Issue



Vertical lines represent 95% confidence intervals.
Standard errors are cluster−robust at the Ad level.
Observations — technology: 32,350, education: 26,135, environment: 24,375, police: 26,458, voting: 25,394

Notes: This figure reports comment and reaction rates, expressed as percentages of total reach, by issue and location. Colors denote issue areas: technology fairness, education reform, environmental justice, criminal justice and police reform, and voting rights. Vertical lines represent 95% confidence intervals, and standard errors are cluster-robust at the advertisement level. Observations are 32,350 for technology, 26,135 for education, 24,375 for environment, 26,458 for police, and 25,394 for voting.

Figure D4: Comment Characteristics by Location



Notes: This figure reports characteristics of direct text comments by location, expressed as percentages of total comments. Panel (a) shows the share of comments classified as conservative in stance, panel (b) the share with negative sentiment, panel (c) the share classified as offensive, and panel (d) the share classified as informative. The analysis is restricted to direct comments containing text and excludes GIFs or images. Vertical lines represent 95% confidence intervals, and standard errors are cluster-robust at the advertisement level. Total observations are 1,393 comments.

Table D1: Length and Depth of the Comment Section

	Dependent variable:				
	Comment Size	Words per Comment	Sentences per Comment	Sentence Length	Comment Depth
	(1)	(2)	(3)	(4)	(5)
Red	−9.350	−1.046	0.128	−0.912*	0.101
	(8.177)	(1.772)	(0.131)	(0.536)	(0.232)
Swing	10.798	2.780	0.271*	−0.597	0.060
	(11.076)	(2.360)	(0.150)	(0.554)	(0.169)
Constant	115.705***	24.619***	2.244***	10.765***	0.786***
	(7.132)	(1.526)	(0.109)	(0.463)	(0.119)
Observations	2,593	2,593	2,593	2,593	1,393
R^2	0.002	0.002	0.001	0.001	0.0001

Notes: Heteroskedasticity-robust standard errors (HC1) are reported. The unit of observation is the comment. Comment Size is the number of characters excluding spaces. Words per Comment is the number of words. Sentences per Comment is the number of sentences. Sentence Length is token count divided by sentence count. Comment Depth is the number of direct replies to a parent comment, and uses parent comments only. Omitted baseline category is BLUE. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table D2: Quality and Diversity of the Comment Section

	Dependent variable:				
	Reciprocity	Justification	Respect	Simpson's Index	Variance in Political Stance
	(1)	(2)	(3)	(4)	(5)
Red	−0.0003	0.035	−0.112**	−0.049	−0.513***
	(0.049)	(0.047)	(0.049)	(0.038)	(0.196)
Swing	−0.057	0.001	−0.144***	−0.030	−0.260
	(0.046)	(0.039)	(0.048)	(0.038)	(0.194)
Constant	0.331***	0.182***	0.528***	0.567***	1.903***
	(0.035)	(0.031)	(0.036)	(0.028)	(0.153)
Observations	145	145	145	134	134
R^2	0.014	0.006	0.067	0.013	0.054

Notes: Heteroskedasticity-robust standard errors (HC1) are reported. The unit of observation is the post. Reciprocity, Justification, and Respect are binary indicators coded for each direct comment. Each measure is averaged across the direct comments of a post with equal weight. Reciprocity captures whether participants stay on topic and engage with prior comments rather than shifting to unrelated themes, making personal attacks, or using rhetorical questions that hinder argumentation. Justification captures whether participants who express or defend a viewpoint provide reasons or arguments for their position. Respect captures whether participants interact civilly, without threats, insults, vulgar language, identity attacks, humiliating language, or silencing expressions. For the Simpson's Index and Variance in Political Stance, posts with only one usable comment are excluded. Simpson's Index is computed as $1 - \sum_{k=1}^{K} p_k^2$, where p_k denotes the share of comments in political-stance category k within a post. Political Stance is coded on a five-point scale: (1) strongly progressive or left-leaning, (2) slightly or moderately progressive, (3) centrist, unclear, or no explicit stance, (4) slightly or moderately conservative, and (5) strongly conservative or right-leaning. Omitted baseline category is BLUE. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

E Field Experiment: Additional Results and Robustness

Figure E1: Example of Opposing Comment and Hide Option



Notes: This figure presents an example comment and the hide option on Facebook.

Table E1: Comments Displayed in the Field Experiment

Group	Comment (verbatim)	St	Se	In	Of
Supportive comments					
1	They also bulldozed black communities to build sports stadiums and other structures.	1	3	2	1
1	The way that caucasian people study and stalk anything to do with black people just so they can be hateful, hurtful, and racist is a mental illness that needs to be studied, it really is.	1	2	1	2
2	Yep put polluting plants upstream so the water gets polluted, then allow poor people to live nearby for work	1	4	2	1
2	Environmental racism	1	4	1	1
3	Because unfortunately still in 2025 we are racially bias. And anyone who says that ain't true lives in a world all their own.	1	4	1	1
3	The asshats laughing have a huge lack of morals.	1	3	1	2
Mixed comments					
1	It's simple data collection, one thing America excels at. The locations industry has placed factory farms (concentration camps for non-human beings)slaughterhouses, rendering plants, are always in depressed areas where poc live and are perversely affected. You'd never see a pig farm or slaughterhouse placed in suburban or high income areas Oh no! Can't remind the mindless what their violent diets require. They might regain their long lost conscience and feel disgusted at how highly intelligent, sentient animals are violated and tortured. They'd gave to smell death.	1	4	2	2
1	this is the stupidest thing I have seen on here all month, and that's saying a lot!!	5	1	1	2
2	No green new scam wasting millions let people choose their car	5	1	1	1
2	Racism, (im)pure and simple.	1	5	1	1
3	I THINK MOST WHITE FOLKS DON'T CARE WHAT HAPPENS TO THE BLACK PEOPLE AS THEY ALWAYS HAVE.	1	2	1	2
3	This is the stupidest thing I read today. Black communities do not have more pollution or any different water than anyone else!	5	1	1	1
Opposing comments					
1	Nonsense. Places ran by democrats are more likely to face pollution risks	5	1	1	1
1	What will be next, no end to the b.s. .	5	1	1	1
2	Soooooooooooooooooooooooo, WOKE & more DEI, B/S!!!! Thank God we're moving back to "Common Sense"!	5	1	1	2
2	Typical Progressive racism claptrap.	5	1	1	2
3	it depends on the politicians you elect in those cities. Has nothing to do with the color of your skin.	4	2	2	1
3	Enough with the race-baiting bullshit	5	1	1	2

Notes: This table reports the exact text of all comments displayed in the field experiment (Section 5), with their GPT-4 scores. Commenter names and profile images are withheld. Each treatment post displayed two comments: Supportive posts two progressive comments, Opposing posts two conservative comments, and Mixed one of each; the three groups (Grp) correspond to the post triplets matched on reactions. Columns: St = Political Stance (1 = strongly progressive ... 5 = strongly conservative); Se = Sentiment toward the ad (1 = highly negative ... 5 = very positive); In = Informativeness (1 = not ... 3 = highly informative); Of = Offensiveness (1 = not ... 3 = highly offensive). Emoji have been removed for typesetting.

Table E2: Summary Statistics

Variable name	Mean (%)	St. Dev.	US Meta users 18+ (%)
Treatment assignment (obs.)			
Arm: Control (no comments)	25.019	43.312	–
Arm: Supportive	24.881	43.232	–
Arm: Mixed	24.971	43.284	–
Arm: Opposing	25.129	43.376	–
Main Outcomes			
All engagement	0.535	7.292	–
Post expansions	0.290	5.380	–
Interactions	0.022	1.487	–
Link clicks	0.240	4.889	–
Page views	0.183	4.276	–
Demographics			
Gender: females	47.672	49.946	53.3
Gender: males	52.328	49.946	46.8
Age: 18–24	8.692	28.172	18.0
Age: 25–34	30.165	45.897	24.4
Age: 35–44	27.158	44.478	19.2
Age: 45–54	16.262	36.902	14.0
Age: 55–64	10.241	30.319	11.8
Age: 65+	7.482	26.309	12.7
Observations:	**1,054,015**		

Notes: Values are expressed as percentages of reach. US Meta user data are from DataReportal (Kemp, 2025).

Table E3: Balance Checks: Ad-level Delivery Metrics

	Group Mean / (SD)				*t*–test *p*–value		
Variable	(1) Control	(2) Supportive	(3) Mixed	(4) Opposing	(1)–(2)	(1)–(3)	(1)–(4)
Total Spend	150.763 (1.028)	150.276 (1.202)	150.395 (0.722)	150.542 (1.078)	0.196	0.218	0.531
CPM	9.594 (2.130)	9.639 (2.075)	9.626 (2.065)	9.548 (2.046)	0.949	0.964	0.947
Frequency	1.123 (0.026)	1.118 (0.020)	1.116 (0.023)	1.118 (0.023)	0.502	0.356	0.520
Reach	14650.333 (3234.248)	14569.222 (3150.366)	14622.056 (3136.580)	14714.778 (3103.458)	0.939	0.979	0.952
Spend per User	0.011 (0.003)	0.011 (0.002)	0.011 (0.002)	0.011 (0.002)	0.994	0.952	0.894
Observations	18	18	18	18			

Notes: Each observation is an ad. The p-values are based on t-tests using heteroskedasticity-robust standard errors

Table E4: Audience Saturation under Alternative Audience Size and Reach Estimates

		Campaign Reach Estimate	
		Lower Bound	Upper Bound
Audience Size Estimate		993,572	1,054,015
Lower Bound	904,700	1.098	1.165
Midpoint	975,000	1.019	1.081
Upper Bound	1,045,300	0.950	1.008

Notes: Audience size estimates correspond to Meta's "Estimated Audience Size," defined as the number of accounts meeting the specified targeting criteria. We record these estimates at the start of the campaign. Meta reports these values as potential reach ranges based on recent platform activity and targeting configuration. Estimates may fluctuate over time as platform usage and measurement systems evolve. Campaign reach estimates are measured either at the campaign level (lower bound) or at the ad level and then aggregated across ads (upper bound). Saturation is computed as the reach estimate divided by the corresponding audience estimate.

Table E5: The Impact of the Comment Section on On-platform User Engagement

	Dependent variable:							
	All Engagement		Post Expansions		Interactions		Link Clicks	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Any comments	0.065***		0.049***		0.003		0.016**	
	(0.015)		(0.013)		(0.003)		(0.008)	
	[0.000]		[0.005]		[0.405]		[0.070]	
Opposing		0.087***		0.048***		0.009**		0.034***
		(0.021)		(0.017)		(0.004)		(0.011)
		[0.002]		[0.033]		[0.077]		[0.019]
Mixed		0.062***		0.049***		0.002		0.012
		(0.017)		(0.013)		(0.003)		(0.011)
		[0.003]		[0.002]		[0.531]		[0.399]
Supportive		0.047***		0.049***		−0.003		0.003
		(0.018)		(0.015)		(0.003)		(0.008)
		[0.024]		[0.012]		[0.366]		[0.764]
Constant	0.538***	0.539***	0.158**	0.158**	0.004	0.005	0.379***	0.380***
	(0.120)	(0.115)	(0.063)	(0.063)	(0.010)	(0.010)	(0.070)	(0.066)
	[0.030]	[0.047]	[0.063]	[0.065]	[0.692]	[0.669]	[0.036]	[0.049]
ZIP Code Set FEs	✓	✓	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓	✓	✓
ZIP Code Set × Controls	✓	✓	✓	✓	✓	✓	✓	✓
Mean Y in Control	0.485	0.485	0.254	0.254	0.020	0.020	0.228	0.228
p(Support vs. Oppose)		0.059		0.974		0.002		0.008
Wild p(Support vs. Oppose)		0.140		0.979		0.014		0.034
p(Oppose vs. Mixed)		0.216		0.935		0.091		0.106
Wild p(Oppose vs. Mixed)		0.333		0.944		0.167		0.193
p(Support vs. Mixed)		0.394		0.960		0.078		0.432
Wild p(Support vs. Mixed)		0.458		0.967		0.126		0.539
Observations	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015
R^2	0.0004	0.0004	0.0005	0.0005	0.0003	0.0003	0.0002	0.0002

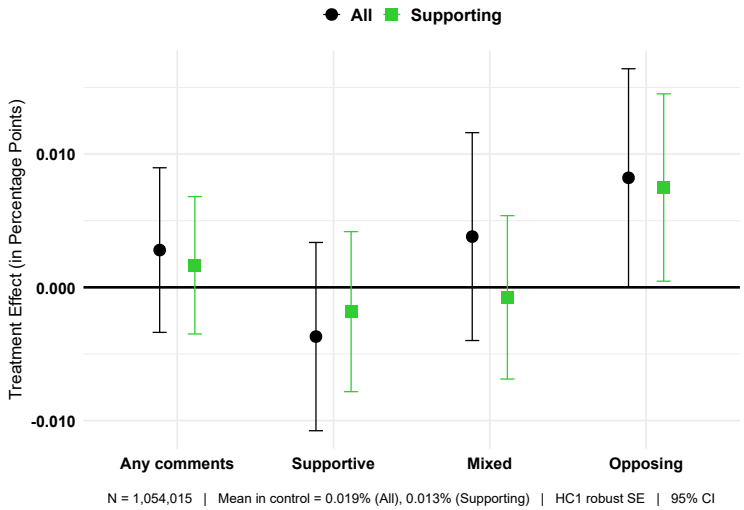
Notes: Standard errors clustered at the advertisement level (72 ads) are reported in parentheses; wild cluster bootstrap p-values are reported in square brackets. Asterisks refer to the cluster-robust standard errors. The unit of observation is the user, constructed from ad-level aggregate data. Controls include user gender and age, as well as their pairwise interactions. The outcome "Interactions" includes comments, reactions, and shares. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table E6: The Impact of the Comment Section on Ad-level Cost Metrics

	Group Mean / (SD)				*t*–test *p*–value		
Variable	(1) Control	(2) Supportive	(3) Mixed	(4) Opposing	(1)–(2)	(1)–(3)	(1)–(4)
Spend per engagement	2.131 (0.337)	1.924 (0.208)	1.879 (0.283)	1.813 (0.392)	0.033	0.021	0.011
Spend per interaction	68.587 (42.266)	88.450 (48.005)	54.910 (29.021)	36.098 (13.111)	0.217	0.288	0.005
Spend per link click	4.685 (1.264)	4.497 (0.817)	4.449 (0.991)	3.996 (0.944)	0.598	0.549	0.066
Observations	18	18	18	18			

Notes: Each observation is an ad. Statistics are weighted by ad reach. The p-values are based on t-tests using heteroskedasticity-robust standard errors. Opposing comments reduce spend per engagement by \$0.30 (15 percent relative to control, $p < 0.05$); mixed and supportive comments also reduce spend per engagement ($p < 0.05$), though by smaller magnitudes. Spend per interaction falls by \$32.5 under opposing comments (a 50 percent reduction, $p < 0.01$), and spend per link click by \$0.69 (15 percent, $p < 0.10$), with no significant effects in the remaining conditions.

Figure E2: The Impact of the Comment Section on All and Supportive Interactions
Outcomes are expressed in % of total reach



Notes: This figure reports treatment effects of comment visibility and stance on all interactions and supportive interactions, expressed in percentage points of total reach. Interactions include comments, reactions, and shares. Supportive interactions include supportive comments, supportive reactions (likes, loves, and cares), and shares. Estimates are obtained from specifications using data collected directly from the posts, with heteroskedasticity-robust (HC1) standard errors. Vertical lines represent 95% confidence intervals. The sample includes 1,054,015 reached users.

Table E7: The Impact of the Comment Section on the Valence of Subsequent Interactions (in %)

	Dependent variable:					
	All		Supportive		Non-supportive	
	(1)	(2)	(3)	(4)	(5)	(6)
Any comments	0.0028		0.0017		0.0011	
	(0.0032)		(0.0026)		(0.0017)	
Opposing		0.0082**		0.0075**		0.0007
		(0.0042)		(0.0036)		(0.0021)
Mixed		0.0038		−0.0008		0.0046*
		(0.0040)		(0.0031)		(0.0025)
Supportive		−0.0037		−0.0018		−0.0019
		(0.0036)		(0.0031)		(0.0019)
Constant	0.0210***	0.0210***	0.0143***	0.0143***	0.0067***	0.0067***
	(0.0034)	(0.0034)	(0.0027)	(0.0027)	(0.0019)	(0.0019)
ZIP Code Set FEs	✓	✓	✓	✓	✓	✓
Mean Y in Control (C)	0.0190	0.0190	0.0133	0.0133	0.0057	0.0057
p(Support vs. Oppose)		0.003		0.008		0.186
p(Oppose vs. Mixed)		0.310		0.020		0.128
p(Support vs. Mixed)		0.048		0.722		0.005
Observations	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015
R^2	0.00001	0.00002	0.000005	0.00001	0.000002	0.00001

Notes: Heteroskedasticity-robust standard errors (HC1) are reported. The unit of observation is the user, constructed from ad-level aggregate data. Supportive interactions include supportive comments, supportive reactions, and shares. Because the stance of comments and the types of reactions are not available from the Meta Ads Manager but are instead retrieved directly from the posts, we cannot include user-level controls and cluster the standard errors at the ad level in this analysis. The outcome "Interactions" includes comments, reactions, and shares.$^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table E8: The Impact of the Comment Section in Blue Areas

	Dependent variable:							
	All Engagement		Post Expansions		Interactions		Link Clicks	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Any comments	0.031		0.014		0.003		0.009	
	(0.025)		(0.024)		(0.005)		(0.015)	
	[0.309]		[0.651]		[0.651]		[0.589]	
Opposing		0.053*		0.023		0.009		0.017
		(0.030)		(0.028)		(0.007)		(0.021)
		[0.199]		[0.530]		[0.293]		[0.542]
Mixed		0.053*		0.029		0.002		0.022
		(0.032)		(0.026)		(0.006)		(0.025)
		[0.234]		[0.381]		[0.800]		[0.548]
Supportive		−0.014		−0.010		−0.003		−0.012
		(0.031)		(0.028)		(0.005)		(0.015)
		[0.714]		[0.792]		[0.640]		[0.486]
Constant	0.559***	0.561***	0.155**	0.156**	−0.007	−0.007	0.416***	0.417***
	(0.129)	(0.117)	(0.072)	(0.066)	(0.013)	(0.012)	(0.077)	(0.072)
	[0.028]	[0.035]	[0.085]	[0.060]	[0.588]	[0.608]	[0.019]	[0.026]
ZIP Code Set FEs	✓	✓	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓	✓	✓
ZIP Code Set × Controls	✓	✓	✓	✓	✓	✓	✓	✓
Mean Y in Control	0.529	0.529	0.267	0.267	0.020	0.020	0.267	0.267
p(Support vs. Oppose)		0.052		0.191		0.029		0.195
Wild p(Support vs. Oppose)		0.148		0.319		0.093		0.334
p(Oppose vs. Mixed)		0.996		0.788		0.270		0.862
Wild p(Oppose vs. Mixed)		0.997		0.823		0.405		0.914
p(Support vs. Mixed)		0.065		0.074		0.363		0.184
Wild p(Support vs. Mixed)		0.202		0.182		0.465		0.376
Observations	329,844	329,844	329,844	329,844	329,844	329,844	329,844	329,844
R^2	0.0004	0.0004	0.0005	0.0005	0.0003	0.0003	0.0002	0.0002

Notes: Standard errors clustered at the advertisement level (24 ads) are reported in parentheses; wild cluster bootstrap p-values are reported in square brackets. Asterisks refer to the cluster-robust standard errors. The unit of observation is the user, constructed from ad-level aggregate data. Controls include user gender and age, as well as their pairwise interactions. The outcome "Interactions" includes comments, reactions, and shares. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table E9: The Impact of the Comment Section in Swing Areas

	Dependent variable:							
	All Engagement		Post Expansions		Interactions		Link Clicks	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Any comments	0.064***		0.044***		0.000		0.025**	
	(0.020)		(0.014)		(0.005)		(0.011)	
	[0.016]		[0.027]		[0.954]		[0.085]	
Opposing		0.071*		0.040*		0.001		0.037**
		(0.037)		(0.025)		(0.008)		(0.015)
		[0.201]		[0.268]		[0.904]		[0.101]
Mixed		0.052***		0.039***		0.004		0.017
		(0.019)		(0.014)		(0.005)		(0.014)
		[0.043]		[0.041]		[0.449]		[0.331]
Supportive		0.070***		0.052***		−0.004		0.020
		(0.022)		(0.015)		(0.006)		(0.013)
		[0.023]		[0.028]		[0.563]		[0.207]
Constant	0.317***	0.317***	0.181***	0.181***	0.021	0.021	0.110**	0.109**
	(0.067)	(0.068)	(0.047)	(0.047)	(0.017)	(0.017)	(0.051)	(0.053)
	[0.028]	[0.032]	[0.007]	[0.004]	[0.344]	[0.389]	[0.089]	[0.092]
ZIP Code Set FEs	✓	✓	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓	✓	✓
ZIP Code Set × Controls	✓	✓	✓	✓	✓	✓	✓	✓
Mean Y in Control	0.478	0.478	0.254	0.254	0.025	0.025	0.209	0.209
p(Support vs. Oppose)		0.996		0.628		0.556		0.282
Wild p(Support vs. Oppose)		0.997		0.737		0.664		0.438
p(Oppose vs. Mixed)		0.604		0.956		0.789		0.237
Wild p(Oppose vs. Mixed)		0.730		0.969		0.839		0.406
p(Support vs. Mixed)		0.382		0.327		0.219		0.844
Wild p(Support vs. Mixed)		0.455		0.392		0.320		0.878
Observations	345,109	345,109	345,109	345,109	345,109	345,109	345,109	345,109
R^2	0.0005	0.0005	0.0006	0.0006	0.0003	0.0003	0.0002	0.0002

Notes: Standard errors clustered at the advertisement level (24 ads) are reported in parentheses; wild cluster bootstrap p-values are reported in square brackets. Asterisks refer to the cluster-robust standard errors. The unit of observation is the user, constructed from ad-level aggregate data. Controls include user gender and age, as well as their pairwise interactions. The outcome "Interactions" includes comments, reactions, and shares. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table E10: The Impact of the Comment Section in Red Areas

	Dependent variable:							
	All Engagement		Post Expansions		Interactions		Link Clicks	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Any comments	0.096***		0.083***		0.004		0.014	
	(0.030)		(0.024)		(0.004)		(0.014)	
	[0.012]		[0.026]		[0.418]		[0.402]	
Opposing		0.132***		0.077**		0.015***		0.045**
		(0.041)		(0.034)		(0.005)		(0.021)
		[0.015]		[0.103]		[0.031]		[0.137]
Mixed		0.077**		0.076***		0.001		−0.003
		(0.031)		(0.024)		(0.005)		(0.016)
		[0.068]		[0.026]		[0.869]		[0.876]
Supportive		0.080**		0.097***		−0.002		−0.000
		(0.031)		(0.027)		(0.004)		(0.014)
		[0.058]		[0.020]		[0.681]		[0.995]
Constant	0.482***	0.482***	0.286**	0.286**	−0.015**	−0.014**	0.212***	0.212***
	(0.152)	(0.145)	(0.121)	(0.120)	(0.006)	(0.007)	(0.049)	(0.041)
	[0.059]	[0.044]	[0.129]	[0.135]	[0.055]	[0.086]	[0.010]	[0.025]
ZIP Code Set FEs	✓	✓	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓	✓	✓
ZIP Code Set × Controls	✓	✓	✓	✓	✓	✓	✓	✓
Mean Y in Control	0.455	0.455	0.242	0.242	0.016	0.016	0.210	0.210
p(Support vs. Oppose)		0.145		0.539		0.000		0.027
Wild p(Support vs. Oppose)		0.290		0.655		0.003		0.113
p(Oppose vs. Mixed)		0.127		0.976		0.001		0.027
Wild p(Oppose vs. Mixed)		0.261		0.981		0.004		0.132
p(Support vs. Mixed)		0.919		0.349		0.370		0.849
Wild p(Support vs. Mixed)		0.924		0.432		0.449		0.876
Observations	379,062	379,062	379,062	379,062	379,062	379,062	379,062	379,062
R^2	0.0005	0.0005	0.0006	0.0006	0.0002	0.0002	0.0001	0.0002

Notes: Standard errors clustered at the advertisement level (24 ads) are reported in parentheses; wild cluster bootstrap p-values are reported in square brackets. Asterisks refer to the cluster-robust standard errors. The unit of observation is the user, constructed from ad-level aggregate data. Controls include user gender and age, as well as their pairwise interactions. The outcome "Interactions" includes comments, reactions, and shares. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table E11: The Impact of Any Comment on On-platform User Engagement

	All Engagement			Post Expansions			Interactions			Link Clicks		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Any comments	0.065***	0.065***	0.066***	0.049***	0.049**	0.049***	0.003	0.003	0.003	0.016**	0.016	0.016
	(0.015)	(0.025)	(0.024)	(0.013)	(0.019)	(0.019)	(0.003)	(0.004)	(0.004)	(0.008)	(0.014)	(0.014)
Constant	0.566***	0.418***	0.485***	0.282***	0.178***	0.254***	0.034***	0.006	0.020***	0.276***	0.243***	0.228***
	(0.046)	(0.035)	(0.021)	(0.045)	(0.025)	(0.016)	(0.008)	(0.004)	(0.003)	(0.021)	(0.025)	(0.012)
ZIP Code Set FEs	✓			✓			✓			✓		
Controls		✓			✓			✓			✓	
Mean Y in Control (C)	0.485	0.485	0.485	0.254	0.254	0.254	0.020	0.020	0.020	0.228	0.228	0.228
Observations						1,054,015						
R^2	0.0001	0.0002	0.00002	0.0001	0.0002	0.00002	0.00002	0.0001	0.00000	0.0001	0.00004	0.00000

Notes: Standard errors are clustered at the advertisement level (72 ads). The unit of observation is the user, constructed from ad-level aggregate data. Columns (1), (4), (7), (10) include ZIP code set fixed effects only. Columns (2), (5), (8), (11) include gender and age controls only. Columns (3), (6), (9), (12) do not include ZIP code set fixed effects and controls. The outcome "Interactions" includes comments, reactions and shares. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table E12: The Impact of Comment Stance on On-platform User Engagement

	All Engagement			Post Expansions			Interactions			Link Clicks		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Opposing	0.086***	0.087***	0.087***	0.048***	0.048*	0.048*	0.009**	0.009*	0.009*	0.033***	0.034**	0.034**
	(0.021)	(0.032)	(0.031)	(0.017)	(0.025)	(0.025)	(0.004)	(0.005)	(0.005)	(0.011)	(0.017)	(0.017)
Mixed	0.061***	0.061**	0.062**	0.049***	0.049**	0.050**	0.002	0.002	0.002	0.011	0.011	0.011
	(0.017)	(0.029)	(0.028)	(0.013)	(0.021)	(0.020)	(0.003)	(0.004)	(0.004)	(0.011)	(0.018)	(0.018)
Supportive	0.047***	0.047*	0.048*	0.049***	0.049**	0.049**	−0.003	−0.003	−0.003	0.002	0.003	0.003
	(0.017)	(0.028)	(0.028)	(0.015)	(0.023)	(0.022)	(0.003)	(0.004)	(0.004)	(0.008)	(0.014)	(0.014)
Constant	0.566***	0.418***	0.485***	0.282***	0.178***	0.254***	0.034***	0.006	0.020***	0.276***	0.243***	0.228***
	(0.046)	(0.035)	(0.021)	(0.045)	(0.025)	(0.016)	(0.006)	(0.004)	(0.003)	(0.021)	(0.025)	(0.012)
ZIP Code Set FEs	✓			✓			✓			✓		
Controls		✓			✓			✓			✓	
Mean Y in Control (C)	0.485	0.485	0.485	0.254	0.254	0.254	0.020	0.020	0.020	0.228	0.228	0.228
p(Support vs. Oppose)	0.067	0.185	0.187	0.935	0.978	0.964	0.003	0.010	0.010	0.007	0.027	0.027
p(Oppose vs. Mixed)	0.230	0.406	0.396	0.914	0.948	0.944	0.101	0.123	0.121	0.105	0.218	0.217
p(Support vs. Mixed)	0.410	0.587	0.591	0.987	0.968	0.982	0.081	0.144	0.139	0.427	0.585	0.590
Observations						1,054,015						
R^2	0.0001	0.0002	0.00002	0.0001	0.0002	0.00002	0.00003	0.0001	0.00001	0.0001	0.00005	0.00001

Notes: Standard errors are clustered at the advertisement level (72 ads). The unit of observation is the user, constructed from ad-level aggregate data. Columns (1), (4), (7), (10) include ZIP code set fixed effects only. Columns (2), (5), (8), (11) include gender and age controls only. Columns (3), (6), (9), (12) do not include ZIP code set fixed effects and controls. The outcome "Interactions" includes comments, reactions and shares. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table E13: The Impact of the Comment Section on On-platform User Engagement (Logistic Regression)

	Dependent variable (Odds Ratios)							
	All Engagement		Post Expansions		Interactions		Link Clicks	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Any comments	1.136***		1.193***		1.131		1.070	
	(0.036)		(0.052)		(0.177)		(0.050)	
Opposing		1.181***		1.191***		1.436**		1.148**
		(0.045)		(0.063)		(0.257)		(0.064)
Mixed		1.128***		1.196***		1.110		1.051
		(0.043)		(0.063)		(0.210)		(0.060)
Supportive		1.098**		1.193***		0.848		1.012
		(0.043)		(0.063)		(0.172)		(0.058)
Constant	0.005***	0.005***	0.002***	0.002***	0.000	0.000	0.004***	0.004***
	(0.001)	(0.001)	(0.001)	(0.001)	(0.00000)	(0.00000)	(0.001)	(0.001)
ZIP Code Set FEs	✓	✓	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓	✓	✓
ZIP Code Set × Controls	✓	✓	✓	✓	✓	✓	✓	✓
Mean Y in Control (C)	0.485	0.485	0.254	0.254	0.020	0.020	0.228	0.228
Observations	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015

Notes: Coefficients reported as odds ratios. The unit of observation is the user, constructed from ad-level aggregate data. Controls include user gender and age, as well as their pairwise interactions. The outcome "Interactions" includes comments, reactions and shares. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table E14: Ad-level Estimates

	Group Mean / (SD)				t–test p–value			
Variable	(1) Control	(2) Supportive	(3) Mixed	(4) Opposing	(1)–(2)	(1)–(3)	(1)–(4)	(2)–(4)
All Engagement	0.485	0.533	0.547	0.572	0.096	0.032	0.008	0.201
	(0.091)	(0.080)	(0.078)	(0.098)				
Post expansions	0.254	0.303	0.303	0.302	0.036	0.017	0.063	0.965
	(0.073)	(0.065)	(0.046)	(0.080)				
Interactions	0.020	0.017	0.022	0.029	0.506	0.557	0.079	0.014
	(0.014)	(0.014)	(0.009)	(0.015)				
Link Clicks	0.228	0.230	0.239	0.261	0.850	0.543	0.050	0.034
	(0.050)	(0.035)	(0.058)	(0.048)				
Observations	18	18	18	18				

Notes: Each observation corresponds to an ad. Statistics are weighted by ad reach. The p-values are based on t-tests using heteroskedasticity-robust standard errors.

Table E15: The Impact of the Comment Section on On-platform User Engagement (clustered at the ZIP-code-set level)

	Dependent variable:							
	All Engagement		Post Expansions		Interactions		Link Clicks	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Any comments	0.065***		0.049***		0.003		0.016**	
	(0.017)		(0.015)		(0.003)		(0.008)	
	[0.001]		[0.006]		[0.389]		[0.053]	
Opposing		0.087***		0.048**		0.009*		0.034***
		(0.026)		(0.021)		(0.005)		(0.013)
		[0.003]		[0.041]		[0.067]		[0.025]
Mixed		0.062***		0.049***		0.002		0.012
		(0.017)		(0.014)		(0.004)		(0.011)
		[0.001]		[0.004]		[0.575]		[0.300]
Supportive		0.047**		0.049***		−0.003		0.003
		(0.021)		(0.018)		(0.003)		(0.011)
		[0.033]		[0.018]		[0.328]		[0.803]
Constant	0.538***	0.539***	0.158***	0.158***	0.004	0.005	0.379***	0.380***
	(0.017)	(0.017)	(0.016)	(0.016)	(0.005)	(0.005)	(0.013)	(0.013)
	[0.239]	[0.237]	[0.459]	[0.465]	[0.564]	[0.551]	[0.210]	[0.204]
ZIP Code Set FEs	✓	✓	✓	✓	✓	✓	✓	✓
Controls	✓	✓	✓	✓	✓	✓	✓	✓
ZIP Code Set × Controls	✓	✓	✓	✓	✓	✓	✓	✓
Mean Y in Control	0.485	0.485	0.254	0.254	0.020	0.020	0.228	0.228
p(Support vs. Oppose)		0.111		0.979		0.021		0.010
Wild p(Support vs. Oppose)		0.132		0.977		0.029		0.019
p(Oppose vs. Mixed)		0.331		0.945		0.167		0.241
Wild p(Oppose vs. Mixed)		0.347		0.942		0.178		0.256
p(Support vs. Mixed)		0.439		0.959		0.055		0.496
Wild p(Support vs. Mixed)		0.467		0.958		0.067		0.550
Observations	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015	1,054,015
R^2	0.0004	0.0004	0.0005	0.0005	0.0003	0.0003	0.0002	0.0002

Notes: Standard errors clustered at the ZIP-code-set level (18 ZIP-code sets) are reported in parentheses; wild cluster bootstrap p-values are reported in square brackets. Asterisks refer to the cluster-robust standard errors. The unit of observation is the user, constructed from ad-level aggregate data. Controls include user gender and age, as well as their pairwise interactions. The outcome "Interactions" includes comments, reactions, and shares. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table E16: The Impact of the Comment Section on Landing-Page Views

	Dependent variable (as % of total reach):	
	Landing-Page Views	
	(1)	(2)
Any comments	0.016*	
	(0.008)	
Opposing		0.019*
		(0.011)
Mixed		0.017
		(0.011)
Supportive		0.011
		(0.011)
Constant	0.347**	0.347***
	(0.135)	(0.134)
ZIP Code Set FEs	✓	✓
Controls	✓	✓
ZIP Code Set × Controls	✓	✓
Mean Y in Control (C)	0.171	0.171
p(Support vs. Oppose)		0.523
p(Oppose vs. Mixed)		0.880
p(Support vs. Mixed)		0.611
Observations	1,054,015	1,054,015
R^2	0.0002	0.0002

Notes: Standard errors are clustered at the advertisement level (72 ads). The unit of observation is the user, constructed from ad-level aggregate data. Controls include user gender and age, as well as their pairwise interactions. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Table E17: Robustness Check: Excluding Toxic Comments (Threshold = 0.7)

	Dependent variable:			
	All Engagement	Post Expansions	Interactions	Link Clicks
	(1)	(2)	(3)	(4)
Opposing	0.100***	0.059***	0.012***	0.032**
	(0.024)	(0.021)	(0.004)	(0.014)
Mixed	0.085***	0.066***	0.005	0.017
	(0.019)	(0.014)	(0.004)	(0.015)
Supportive	0.047***	0.049***	−0.003	0.002
	(0.016)	(0.015)	(0.003)	(0.008)
Constant	0.521***	0.144**	0.002	0.378***
	(0.111)	(0.061)	(0.010)	(0.066)
ZIP Code Set FEs	✓	✓	✓	✓
Controls	✓	✓	✓	✓
ZIP Code Set × Controls	✓	✓	✓	✓
Mean Y in Control (C)	0.485	0.254	0.020	0.228
p(Support vs. Oppose)	0.024	0.616	0.001	0.039
p(Oppose vs. Mixed)	0.519	0.723	0.164	0.392
p(Support vs. Mixed)	0.044	0.215	0.022	0.332
Observations	878,687	878,687	878,687	878,687
R^2	0.0005	0.001	0.0003	0.0002

Notes: Standard errors are clustered at the advertisement level (60 ads). We excluded posts with any displayed comments that had toxicity scores above 0.7. The unit of observation is the user, constructed from ad-level aggregate data. Controls include user gender and age, as well as their pairwise interactions. The outcome "Interactions" includes comments, reactions and shares. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Figure E3: Sensitivity to Excluding One ZIP-Code Set at a Time: All Engagement

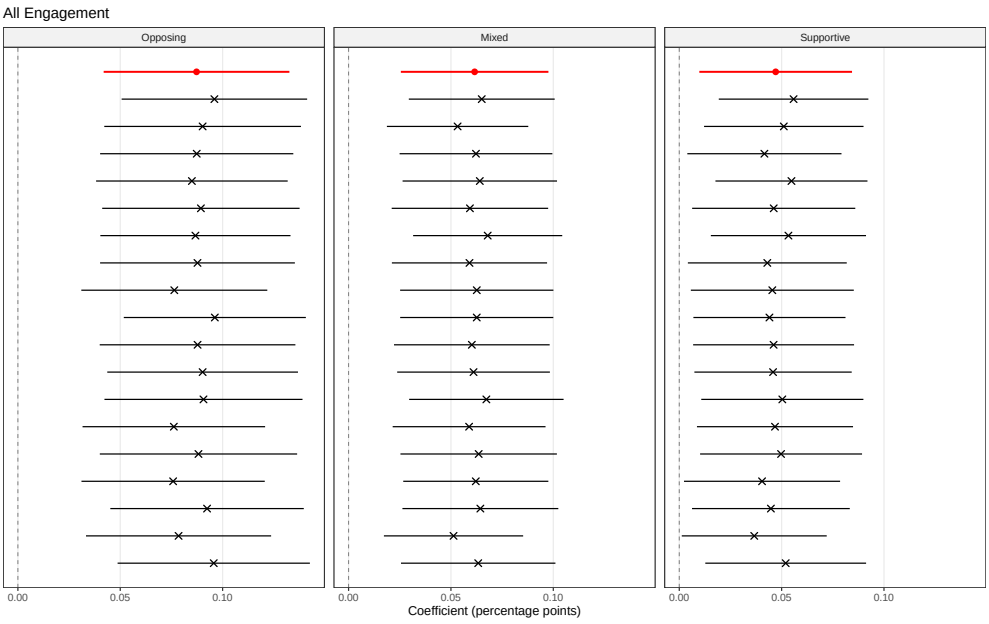


Figure E4: Sensitivity to Excluding One ZIP-Code Set at a Time: Post Expansions

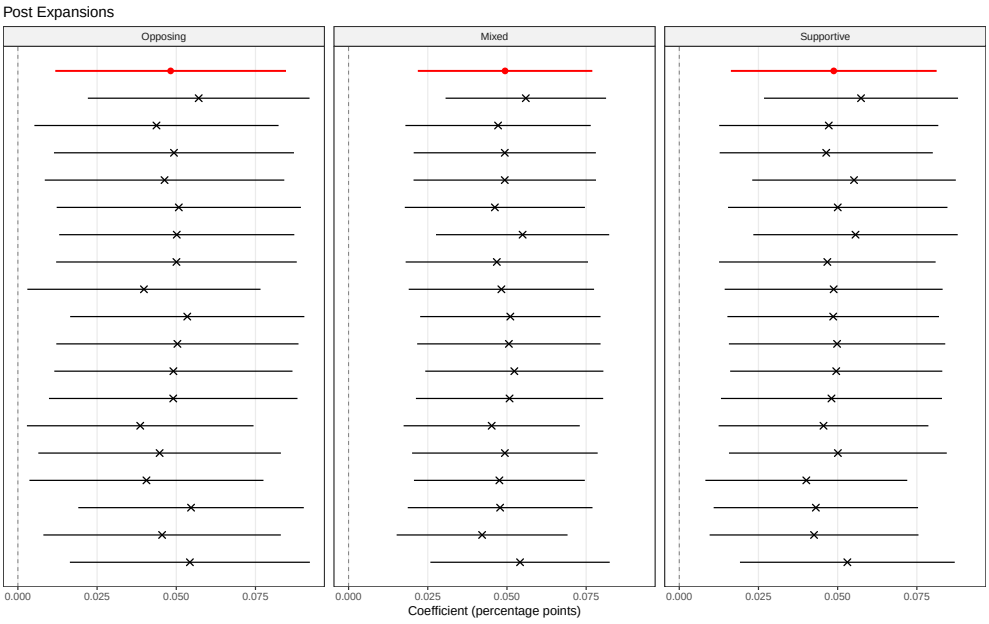


Figure E5: Sensitivity to Excluding One ZIP-Code Set at a Time: Interactions

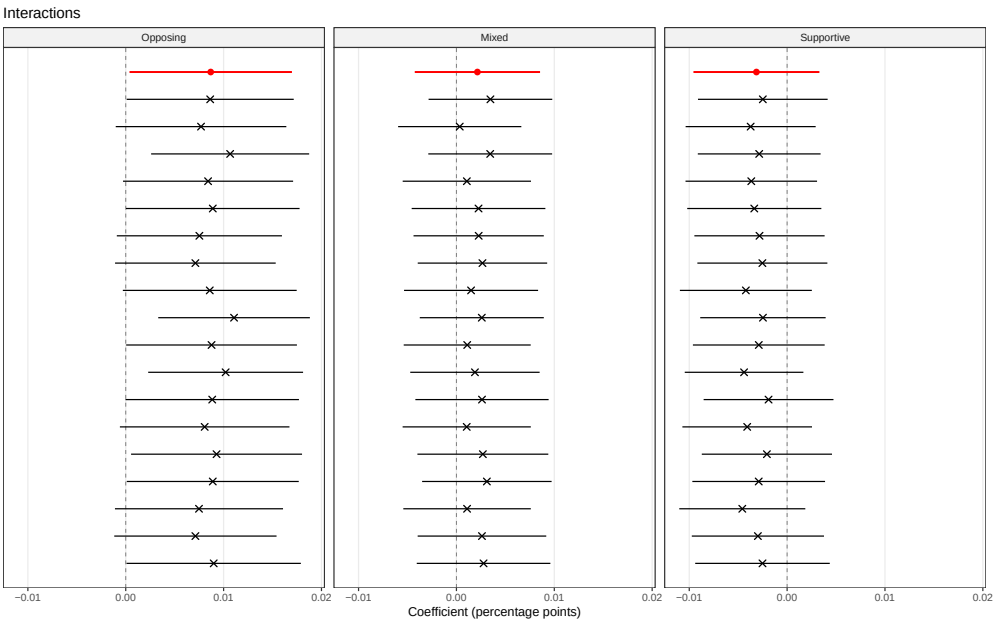
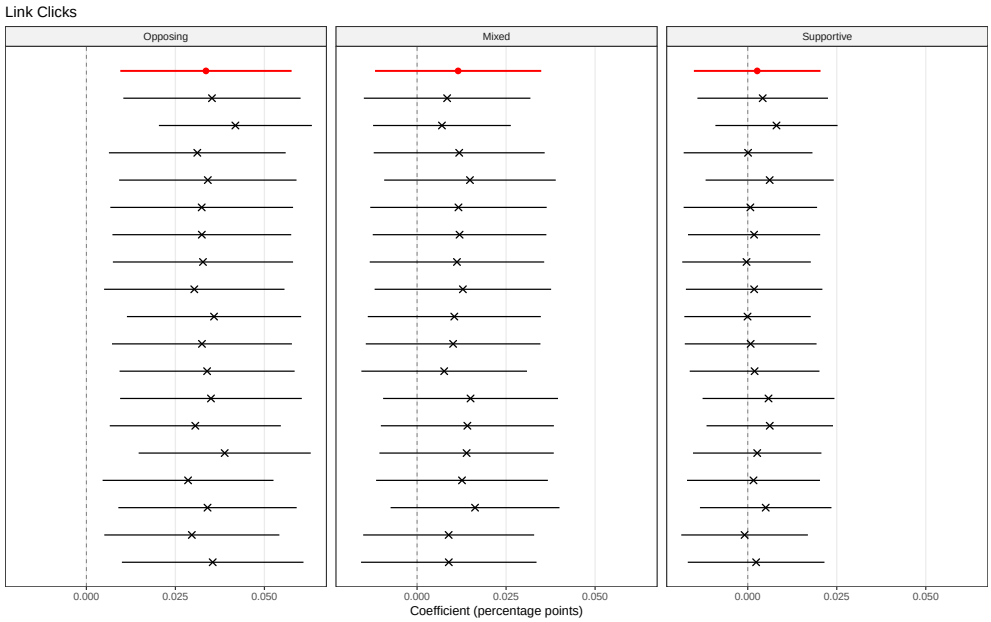


Figure E6: Sensitivity to Excluding One ZIP-Code Set at a Time: Link Clicks



Notes: Figures E3–E6 report a sensitivity analysis in which subsets of ZIP codes are iteratively excluded and the main treatment effects are re-estimated. Each point corresponds to an estimated coefficient from a re-estimated specification, and horizontal lines represent 95% confidence intervals. The red point indicates the baseline estimate from the full sample.

Figure E7: Sensitivity to Excluding One Comment Section at a Time: All Engagement

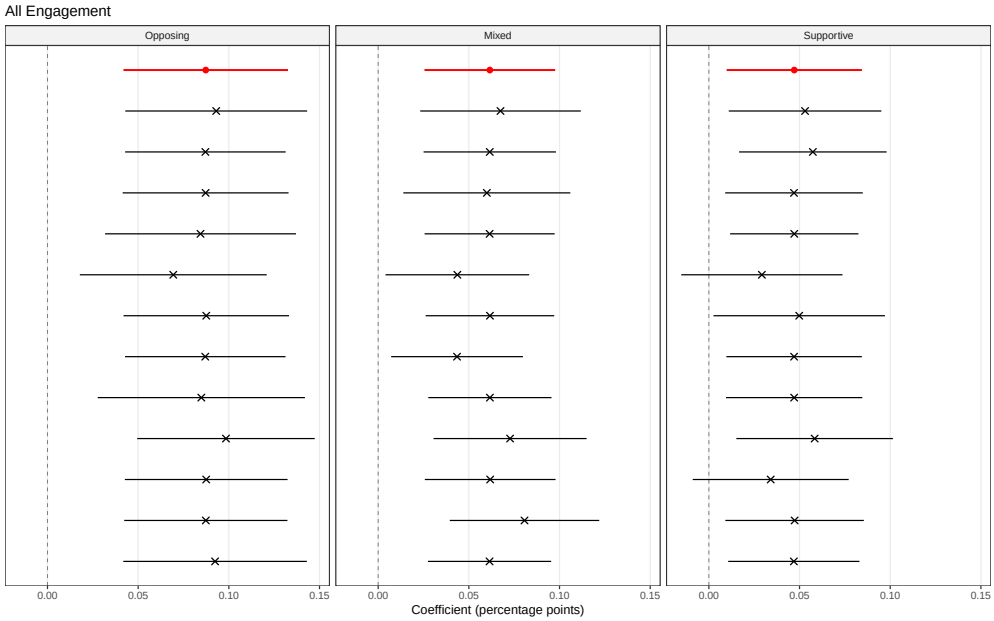


Figure E8: Sensitivity to Excluding One Comment Section at a Time: Post Expansions

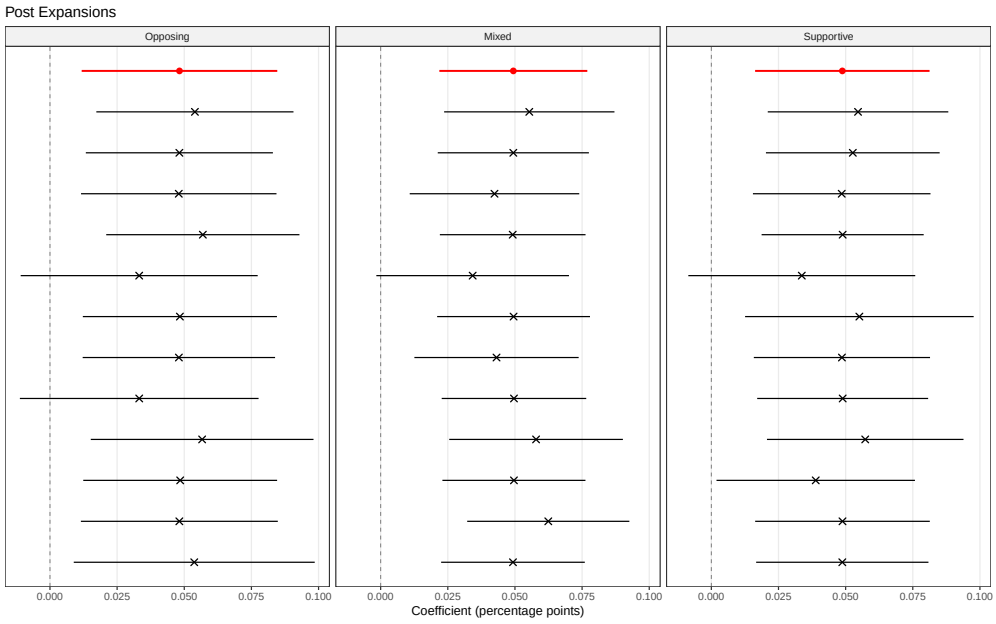


Figure E9: Sensitivity to Excluding One Comment Section at a Time: Interactions

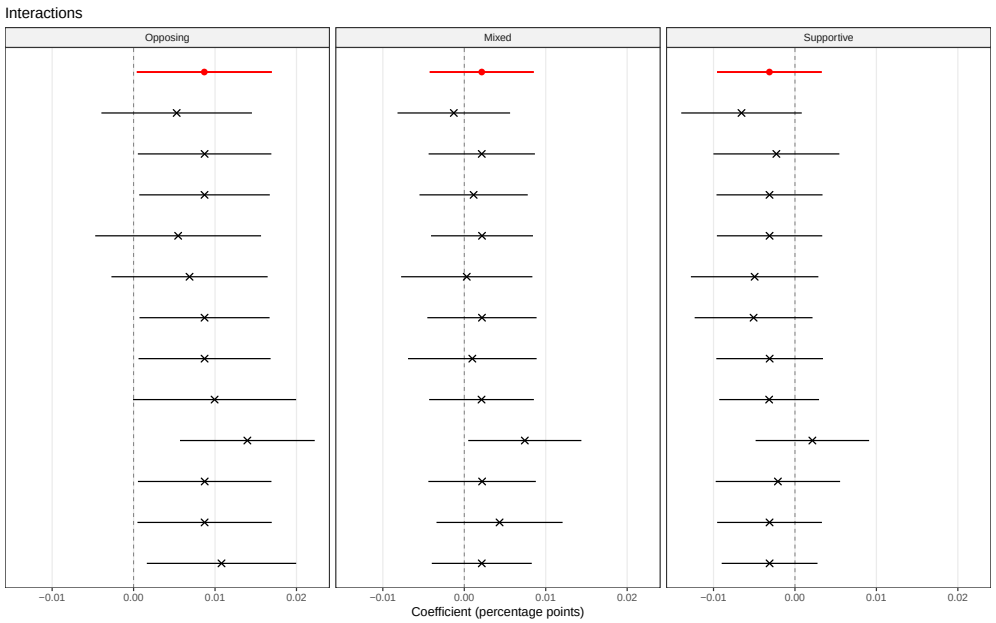
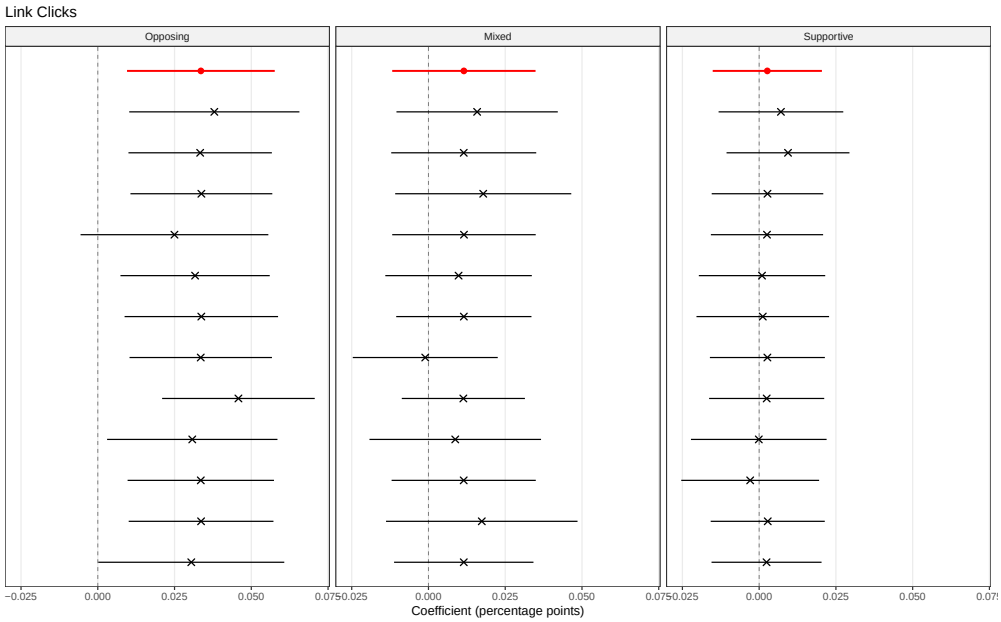


Figure E10: Sensitivity to Excluding One Comment Section at a Time: Link Clicks



Notes: Figures E7–E10 report a sensitivity analysis in which each comment section (post) is iteratively excluded and the main treatment effects are re-estimated. Each point corresponds to an estimated coefficient from a re-estimated specification, and horizontal lines represent 95% confidence intervals. The red point indicates the baseline estimate from the full sample.

Figure E11: Randomization-Inference Tests: All Engagement

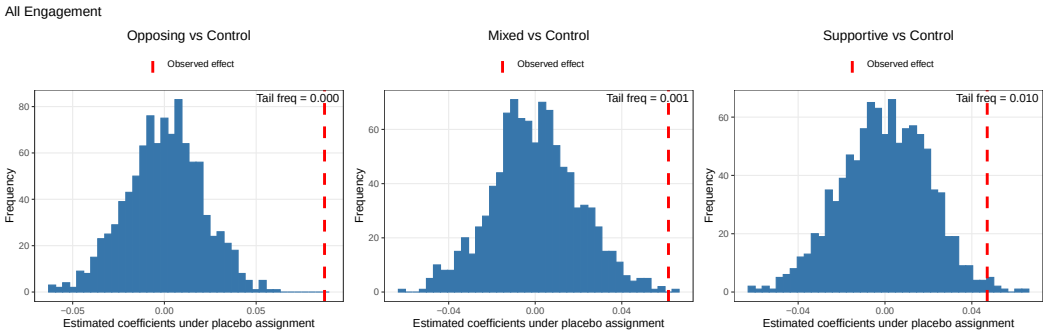


Figure E12: Randomization-Inference Tests: Post Expansions

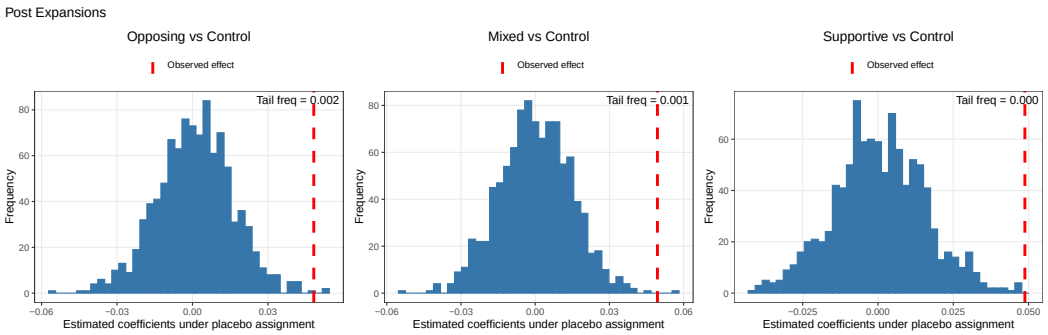


Figure E13: Randomization-Inference Tests: Interactions

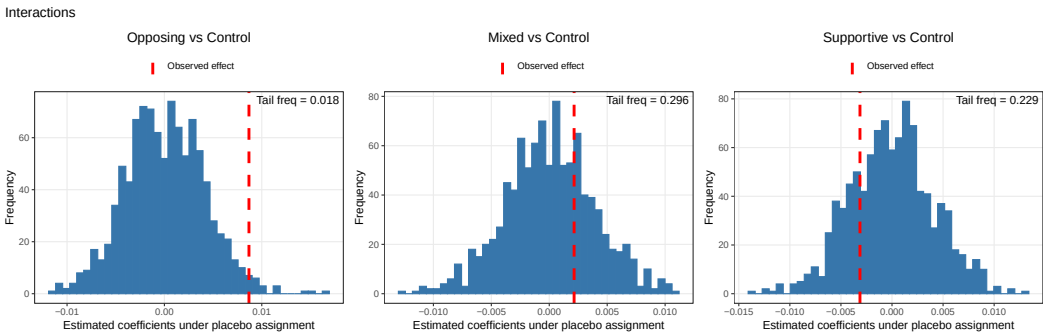
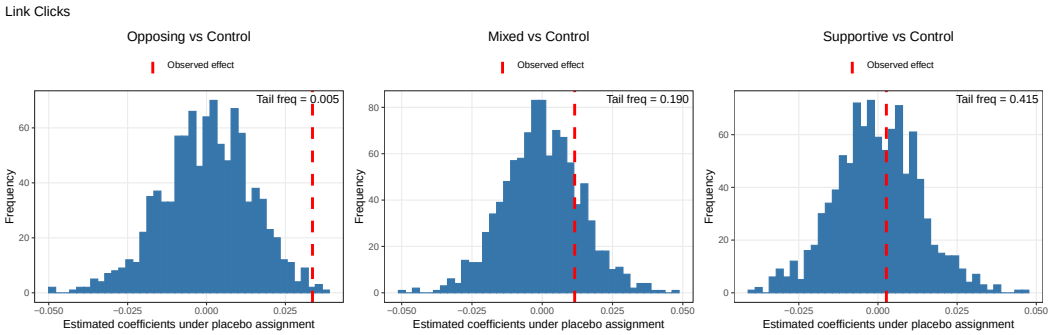


Figure E14: Randomization-Inference Tests: Link Clicks

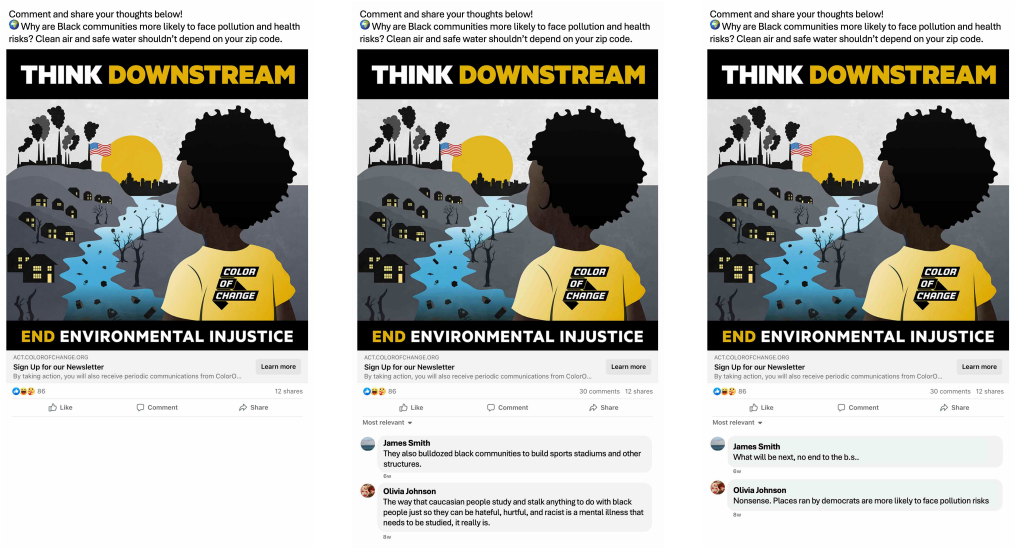


Notes: Figures E11–E14 report randomization-inference placebo tests. In each of 1,000 draws, individuals are randomly reassigned to the four arms within each ZIP-code set, with equal arm sizes, and the full specification of equation (1) is re-estimated. Histograms show the distribution of the resulting placebo estimates. The vertical dashed line marks the observed treatment effect in the data. Tail frequencies report the share of placebo estimates at least as extreme as the observed estimate, in its direction, and therefore correspond to one-sided permutation p-values.

F Survey Experiment

F.1 Design

Figure F1: Survey Experiment Stimuli



(a) Control (b) Supportive (c) Opposing

Notes: This figure displays screenshots of the survey experiment stimuli for the environment issue. Panel (a) shows the Control condition, in which the post is displayed without comments. Panel (b) shows the Supportive condition, in which two supportive comments appear below the post. Panel (c) shows the Opposing condition, in which two opposing comments appear below the post. The environment post is shown for illustration; participants were assigned to view three posts in random order under their assigned comment condition.

Table F1: Sample Demographics Relative to U.S. Adults

	Analysis sample	U.S. adults
Female	0.493	0.505
College (Bachelor's or higher)	0.523	0.362
Age 18–34	0.470	0.373
Age 35–54	0.423	0.420
Age 55–64	0.107	0.206
White	0.791	0.723
Black or African American	0.116	0.144
Asian or Pacific Islander	0.077	0.079
American Indian or Alaska Native	0.023	0.026
Other or mixed race	0.057	0.128
Hispanic or Latino origin	0.107	0.194
Democrat	0.320	0.270
Republican	0.294	0.270
Independent	0.369	0.450
N	3,868	

Notes: This table compares the composition of the analysis sample in the survey experiment with the U.S. adult population. The first column reports shares in the analysis sample, and the second column reports population benchmarks from 2023 American Community Survey (gender, age, race, and Hispanic origin for the total population; college attainment, bachelor's degree or higher, for adults aged 25 and over) and party identification from Gallup 2025. Age shares are computed among adults aged 18–64 to match the survey's inclusion criteria. Race shares use the ACS "race alone or in combination" tabulation and need not sum to one.

Table F2: Survey Completion Rates Across Experimental Arms

Variable	(1) Control Share	(2) Supportive Share	(3) Opposing Share	F-test p-value
Completed Survey	1.00	0.99	0.99	0.134

Notes: This table reports the share of respondents who completed the survey in each experimental arm, among those who were randomized and passed the attention check. The final column reports the p-value from an F-test of joint equality of completion rates across the three arms.

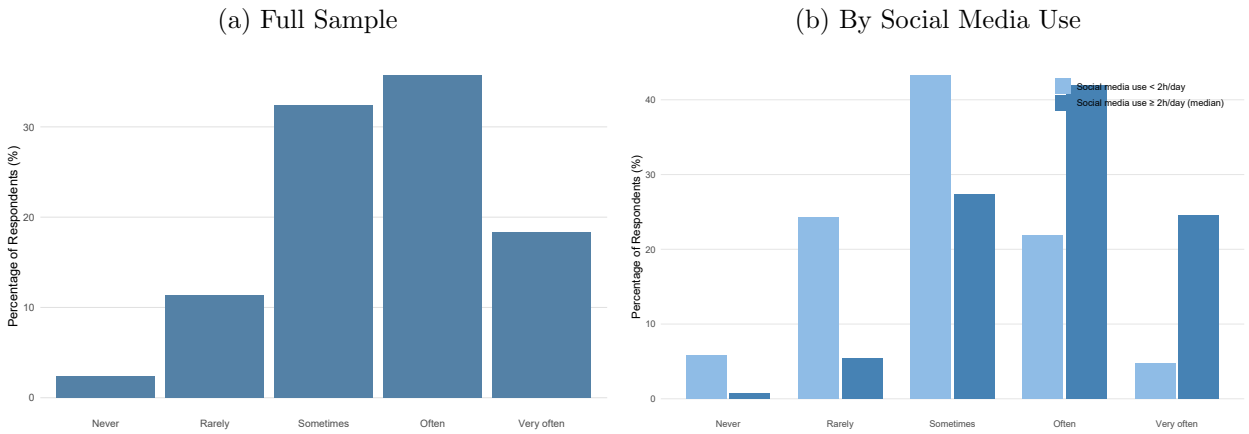
Table F3: Covariate Balance Across Experimental Arms

	(1) Control	(2) Supportive	(3) Opposing	SMD			t-test p-value		
Variable	Mean/SD	Mean/SD	Mean/SD	(2)-(1)	(3)-(1)	(3)-(2)	(1)-(2)	(1)-(3)	(2)-(3)
Age (> Median)	0.48 (0.50)	0.49 (0.50)	0.51 (0.50)	0.025	0.064	0.039	0.54	0.11	0.31
Baseline Racial Index (> Median)	0.49 (0.50)	0.50 (0.50)	0.48 (0.50)	0.013	-0.022	-0.035	0.74	0.58	0.36
Perceived Share of Non-Conservatives (> Median)	0.41 (0.49)	0.38 (0.49)	0.39 (0.49)	-0.072	-0.047	0.025	0.07*	0.24	0.51
Perceived Share of Progressives (> Median)	0.49 (0.50)	0.47 (0.50)	0.47 (0.50)	-0.040	-0.045	-0.005	0.32	0.26	0.91
Education (> Median)	0.16 (0.36)	0.15 (0.36)	0.18 (0.38)	-0.009	0.052	0.062	0.81	0.19	0.11
Political view (> Median)	0.38 (0.49)	0.40 (0.49)	0.39 (0.49)	0.045	0.026	-0.020	0.26	0.51	0.61
Social Media Use (> Median)	0.40 (0.49)	0.41 (0.49)	0.39 (0.49)	0.016	-0.032	-0.049	0.68	0.41	0.21
Reads Comments Frequency (> Median)	0.18 (0.38)	0.19 (0.39)	0.18 (0.39)	0.023	0.003	-0.020	0.57	0.94	0.60
Female	0.51 (0.50)	0.49 (0.50)	0.49 (0.50)	-0.040	-0.037	0.003	0.32	0.35	0.94
White	0.78 (0.41)	0.78 (0.41)	0.81 (0.39)	0.001	0.070	0.069	0.98	0.08*	0.07*
Black / African American	0.12 (0.32)	0.12 (0.33)	0.11 (0.31)	0.015	-0.036	-0.052	0.71	0.36	0.18
Asian American / Pacific Islander	0.08 (0.28)	0.07 (0.26)	0.07 (0.26)	-0.041	-0.043	-0.002	0.31	0.27	0.96
Hispanic	0.09 (0.29)	0.13 (0.33)	0.10 (0.30)	0.104	0.022	-0.082	0.01***	0.59	0.03**
Democrat	0.32 (0.47)	0.31 (0.46)	0.33 (0.47)	-0.007	0.029	0.036	0.87	0.46	0.36
Republican	0.30 (0.46)	0.30 (0.46)	0.28 (0.45)	0.005	-0.036	-0.041	0.91	0.36	0.29
Independent	0.37 (0.48)	0.37 (0.48)	0.37 (0.48)	0.003	-0.004	-0.007	0.94	0.91	0.85
N	1,189	1,287	1,392						
F-test of joint significance (p-value)							0.355	0.527	0.238
F-test, number of observations							2,476	2,581	2,679

Notes: This table reports covariate balance across the three treatment arms of the survey experiment. Columns (1)–(3) report means with standard deviations in parentheses for the control, supportive comments, and opposing comments groups, respectively. The remaining columns report the standardized mean differences (SMD) and p-values from pairwise t-tests of equality of means. The final rows report F-tests of joint significance across all covariates for each pair of arms. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

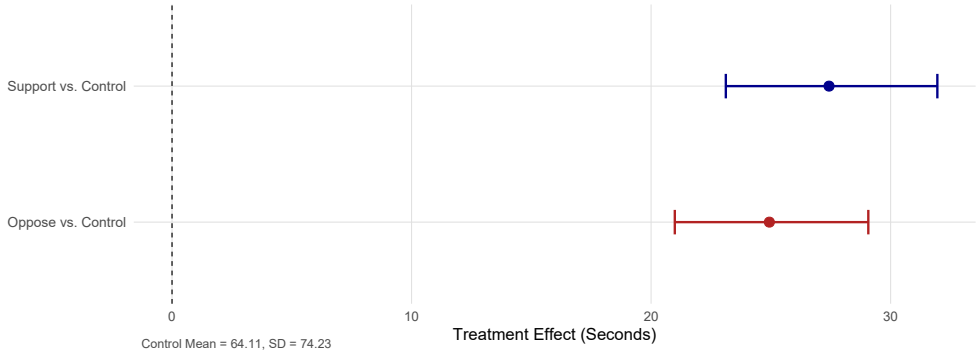
F.2 Additional Results

Figure F2: Reported Frequency of Reading or Checking Comments on Social Media



Notes: This figure reports the self-reported frequency with which survey participants read or check comments on social media. Panel (a) shows the distribution for the full sample. Panel (b) splits the sample by social media use, comparing participants above and below the median level of social media use.

Figure F3: Time Spent on Posts (in seconds)



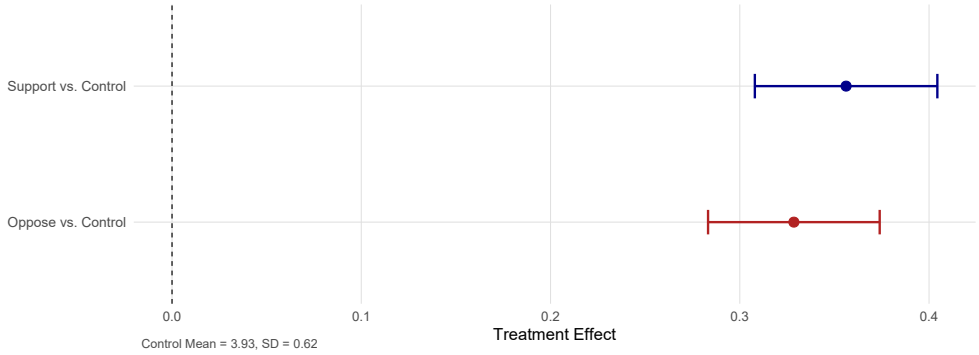
Notes: This figure reports treatment effects of supportive and opposing comments, relative to the no-comments control, on time spent viewing the posts. Effects are expressed in seconds. Horizontal lines represent 95% confidence intervals.

Table F4: Effects of Comment Stance on Time Spent, Anger, Annoyance, and Curiosity

	Time Spent	*Anger/Annoyance*	*Curiosity/Reflection (Incl. Comments)*
	(1)	(2)	(3)
Supportive vs. Control	0.356***		
	(0.025)		
Opposing vs. Control	0.329***		
	(0.023)		
Opposing vs. Supportive	-0.028	0.433***	-0.382***
	(0.023)	(0.036)	(0.035)
Controls	✓	✓	✓
Observations	3,868	2,679	2,679
R^2	0.1293	0.1562	0.1936
Mean Y (Control)	3.931		

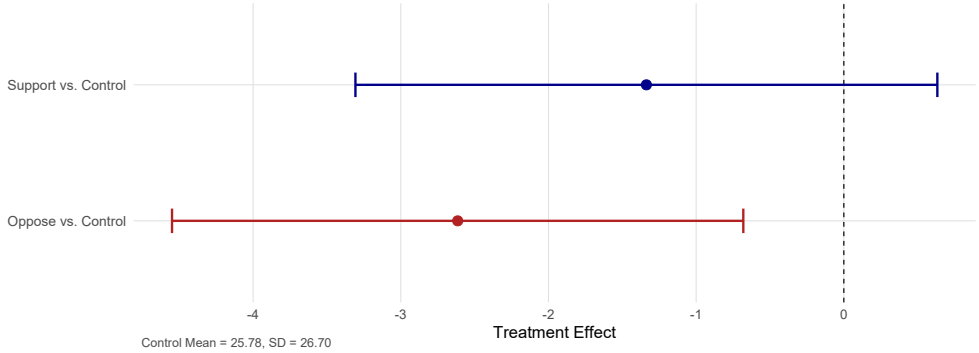
Notes: This table reports OLS estimates from equation (2) for the time spent, anger/annoyance, and curiosity/reflection in the survey experiment. Column 1 shows time spent measured as $\log(1+t)$ where t is total time on the three posts in seconds. Column 2 combines Likert-scale measures of anger and annoyance triggered by the comments, oriented so that higher values indicate a stronger negative emotional response. Because this measure is elicited only from respondents who saw a comment section, it is defined for the respondents in the two treatment arms. Column 3 combines interest in seeing additional comments with how thought-provoking respondents found the post and the comments. Because the comment item is asked only of respondents who saw a comment section, this index, like the anger/annoyance index, is estimated on the respondents in the two treatment arms. The indices are aggregated by inverse-covariance weighting following Anderson (2008). Neither index is observed in the control group, so both are standardized to mean zero and standard deviation one in the pooled sample of the two treatment arms. All columns include the pre-specified baseline controls listed in Section 6, and standard errors robust to heteroskedasticity are in parentheses. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Figure F4: Time Spent on Posts, log(seconds+1)



Notes: This figure reports treatment effects of supportive and opposing comments, relative to the no-comments control, on time spent viewing the posts, expressed in log(seconds+1). Horizontal lines represent 95% confidence intervals.

Figure F5: Donations (U.S. Dollars)



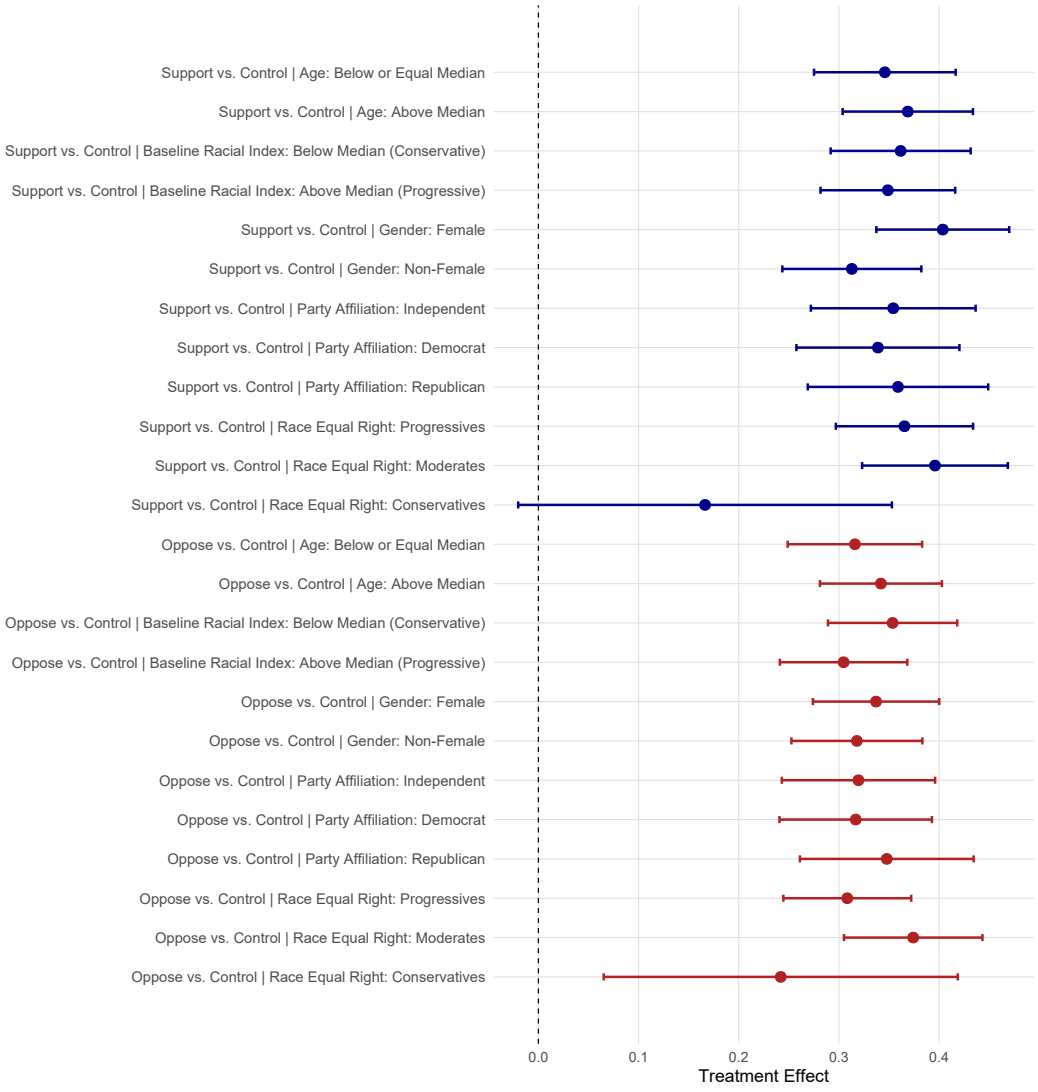
Notes: This figure reports treatment effects of supportive and opposing comments, relative to the control, on the amount donated in the survey experiment, winsorized at the 95th percentile and expressed in U.S. dollars. Horizontal lines represent 95% confidence intervals.

Table F5: Effects of Comment Stance on Attitudes, Donations, and Newsletter Sign-Up

	Post-Exposure Attitudes	*Donated*	*Newsletter Sign-Up*
	(1)	(2)	(3)
Supportive vs. Control	-0.032	-0.028	0.012
	(0.036)	(0.017)	(0.015)
Opposing vs. Control	-0.118***	-0.053***	0.007
	(0.036)	(0.017)	(0.015)
Opposing vs. Supportive	-0.087**	-0.024	-0.006
	(0.035)	(0.017)	(0.015)
Controls	✓	✓	✓
Observations	3,868	3,868	3,868
R^2	0.1885	0.1001	0.1126
Mean Y (Control)	-0.000	0.733	0.193

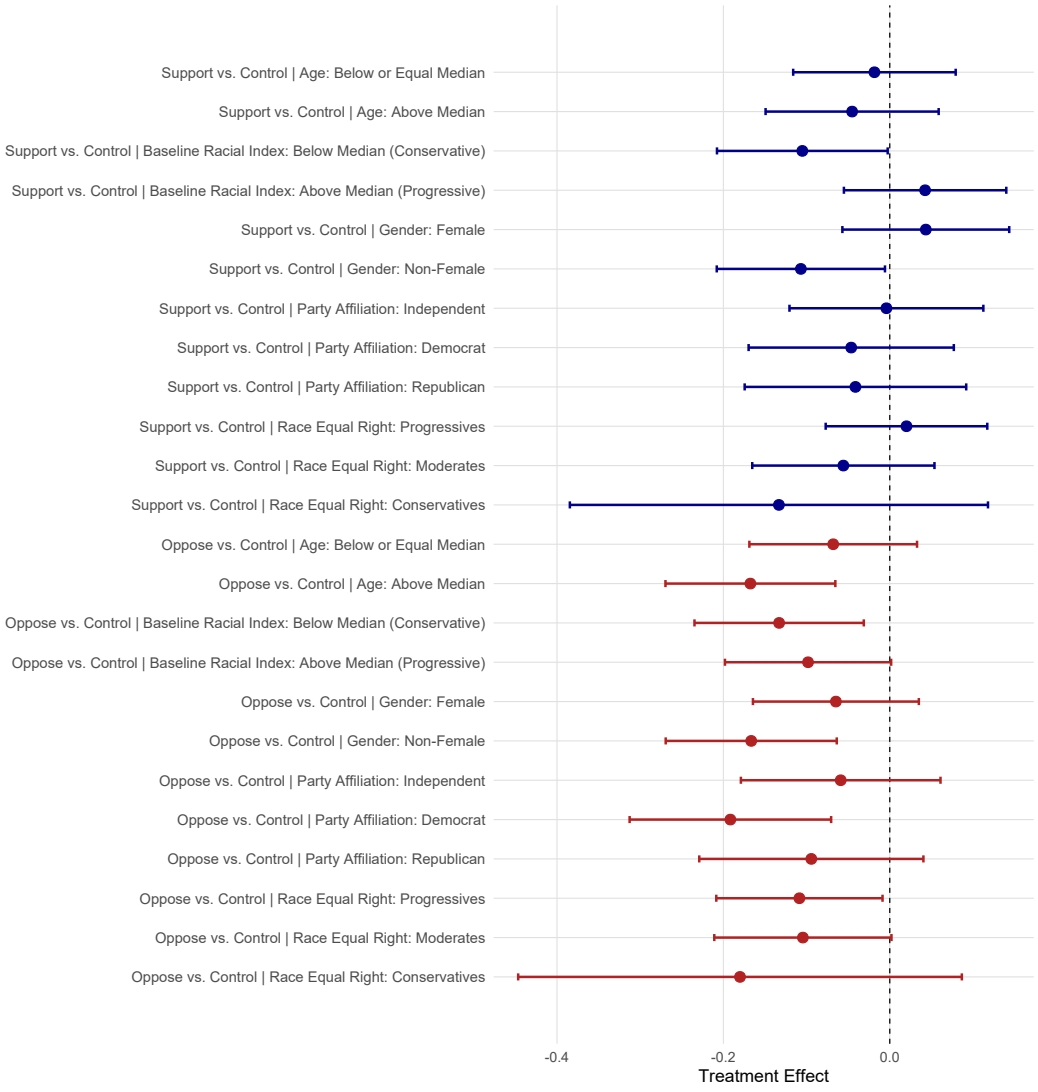
Notes: This table reports OLS estimates from equation (2) for the three primary attitudinal and behavioral outcomes of the survey experiment. *Post-Exposure Attitudes* is the attitudes index described in Section 6, aggregated by inverse-covariance weighting following Anderson (2008) and standardized to mean zero and standard deviation one in the control group, with higher values indicating greater alignment with the organization's position. *Donated* is an indicator for committing a positive amount in the incentivized donation task, and *Newsletter Sign-Up* is an indicator for providing an email address to the organization. All columns include the pre-specified baseline controls listed in Section 6. Standard errors robust to heteroskedasticity are in parentheses. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

Figure F6: Heterogeneous Treatment Effects for Time Spent on the Post, log(seconds+1)



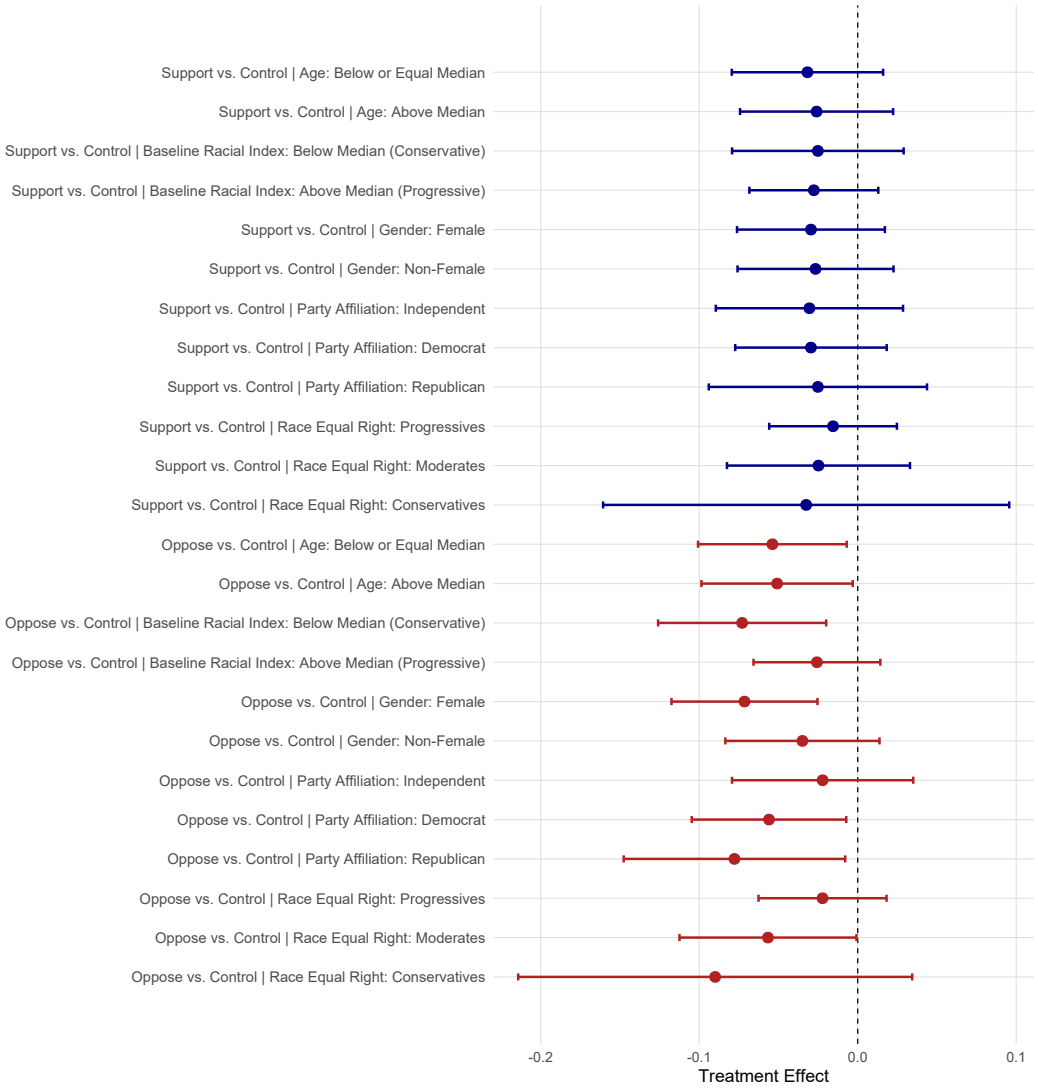
Notes: This figure reports heterogeneous treatment effects in the survey experiment for time spent on the post, expressed in $\log(1+t)$ where t is time in seconds. Effects are shown separately by baseline racial attitudes, party affiliation, gender, age, and ideology. Horizontal lines represent 95% confidence intervals.

Figure F7: Heterogeneous Treatment Effects for Racial Attitudes Index



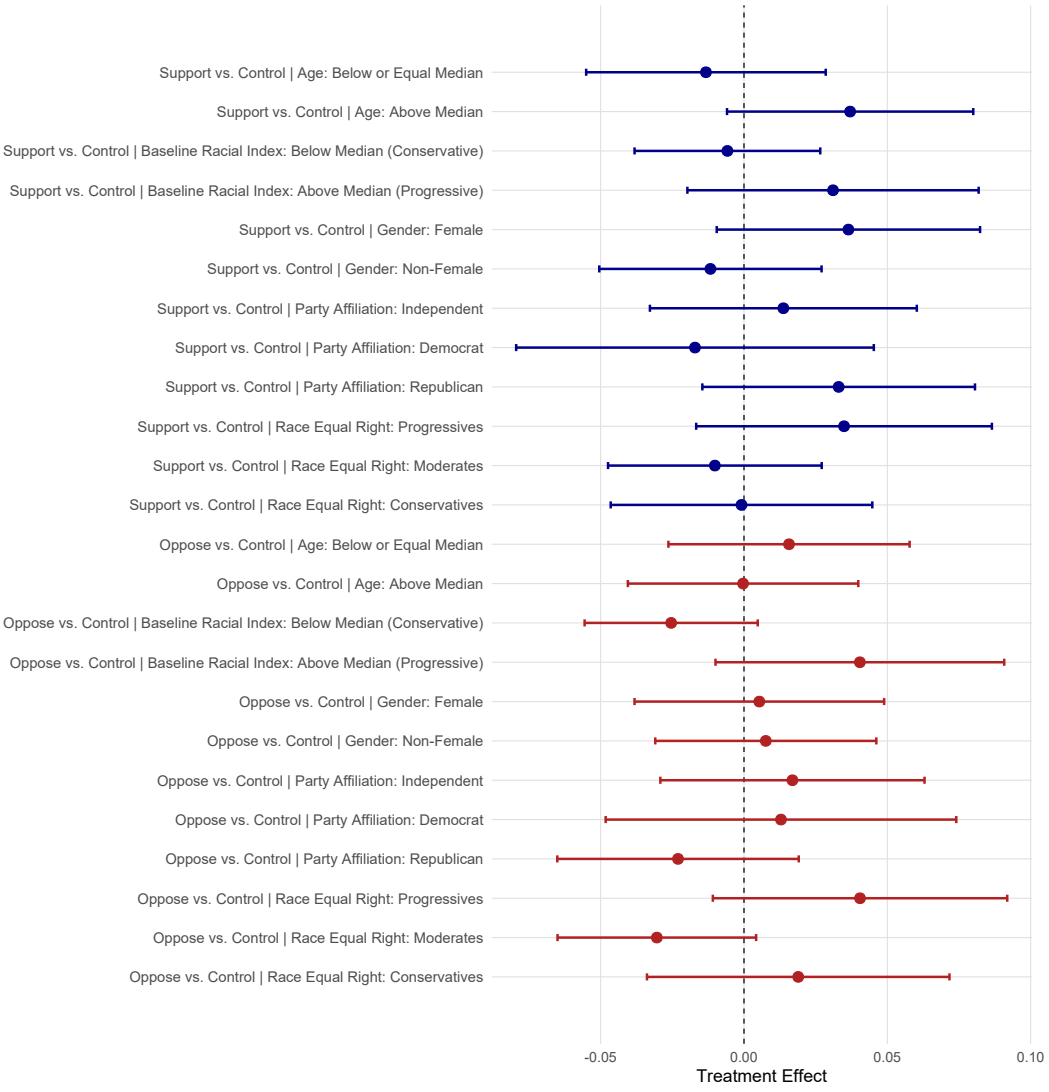
Notes: This figure reports heterogeneous treatment effects in the survey experiment for the racial attitudes index. Effects are shown separately by baseline racial attitudes, party affiliation, gender, age, and ideology. Higher values indicate greater alignment with the organization's position. Horizontal lines represent 95% confidence intervals.

Figure F8: Heterogeneous Treatment Effects for Donation (Yes/No)



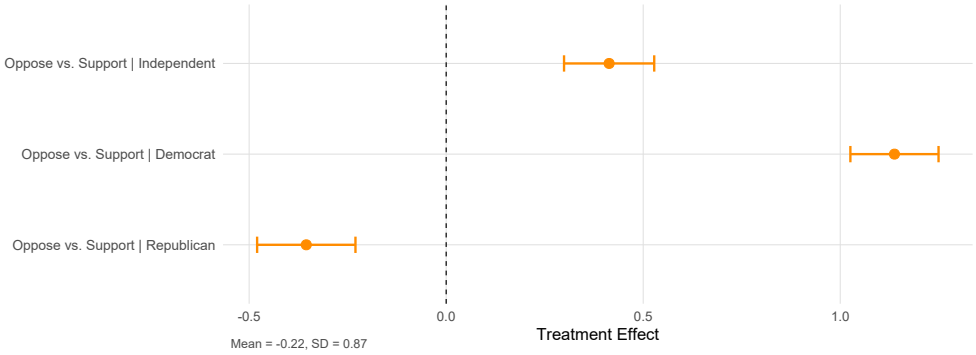
Notes: This figure reports heterogeneous treatment effects in the survey experiment for a binary donation outcome. Effects are shown separately by baseline racial attitudes, party affiliation, gender, age, and ideology. Horizontal lines represent 95% confidence intervals.

Figure F9: Heterogeneous Treatment Effects for Newsletter Sign-Up (Yes/No)



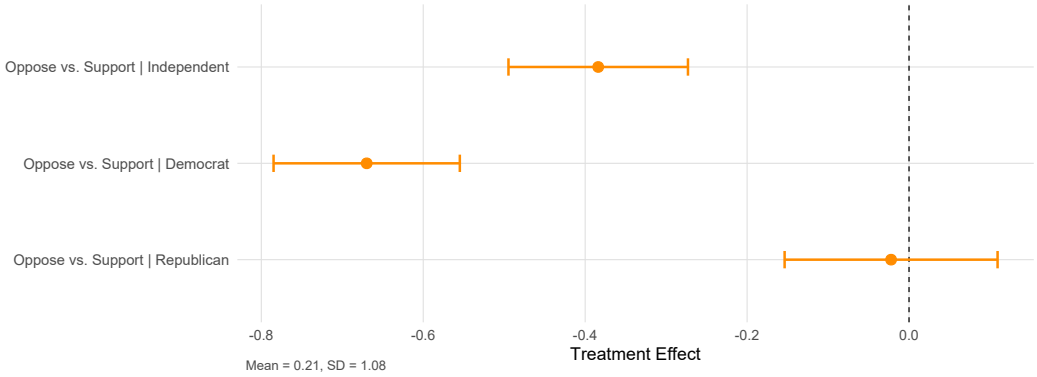
Notes: This figure reports heterogeneous treatment effects in the survey experiment for a binary newsletter sign-up outcome. Effects are shown separately by baseline racial attitudes, party affiliation, gender, age, and ideology. Horizontal lines represent 95% confidence intervals.

Figure F10: Heterogeneous Treatment Effects on Anger and Annoyance by Party Affiliation



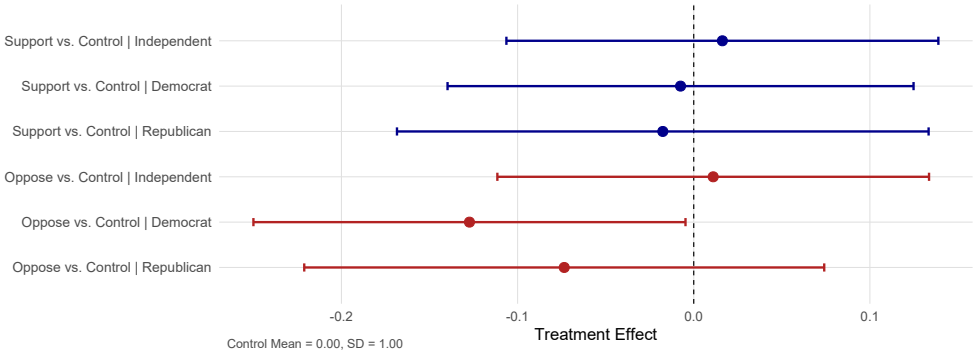
Notes: This figure reports the differential effect of opposing versus supportive comments on an emotional response measure in the survey experiment, separately for Republicans, Democrats, and Independents. The outcome combines self-reported anger and annoyance triggered by the comments, with higher values indicating stronger negative emotional responses. The index is standardized to mean zero and standard deviation one in the pooled sample of the two treatment arms. The mean and standard deviation shown below the figure are for the Supportive arm. Horizontal lines represent 95% confidence intervals.

Figure F11: Heterogeneous Treatment Effects on Curiosity and Reflection by Party Affiliation



Notes: This figure reports the effect of opposing versus supportive comments on an index of curiosity and reflection in the survey experiment, separately for Republicans, Democrats, and Independents. The outcome combines a Likert-scale measure of participants' interest in seeing additional comments on the post (interpreted as curiosity) with measures of how thought-provoking they find the post and the comments, with higher values indicating greater curiosity and reflection. The last item is asked only of respondents who saw comments, so this version of the index is used only to compare the two treatment conditions. The mean and standard deviation shown below the figure are for the Supportive arm. Horizontal lines represent 95% confidence intervals.

Figure F12: Perceived Social Norms



Notes: This figure reports treatment effects of supportive and opposing comments, relative to the control, on perceived social norms in the survey experiment. The outcome is respondents' incentivized estimate of the share of survey participants who agreed or strongly agreed with a conservative statement, reverse-coded so that higher values indicate more progressive perceived norms. Horizontal lines represent 95% confidence intervals.

Table F6: Treatment-Arm Effects Across Alternative Control Sets

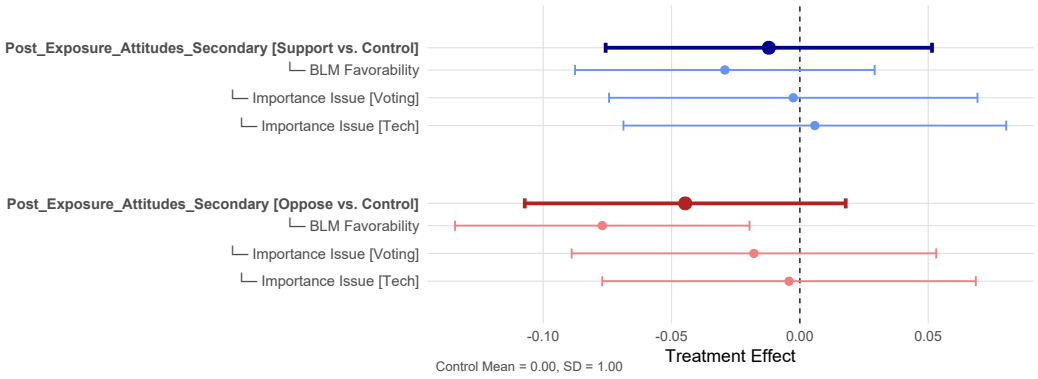
	(1) No Controls	(2) Main	(3) LASSO	(4) Full Controls
A. Post-Exposure Attitudes				
Supportive vs. Control	-0.026	-0.032	-0.034	-0.033
	(0.040)	(0.036)	(0.036)	(0.036)
Opposing vs. Control	-0.125***	-0.118***	-0.118***	-0.121***
	(0.040)	(0.036)	(0.036)	(0.037)
Opposing vs. Supportive	-0.100**	-0.087**	-0.084**	-0.088**
	(0.039)	(0.035)	(0.035)	(0.035)
Observations	3,868	3,868	3,868	3,868
R^2	0.0030	0.1885	0.1914	0.1958
B. Donated				
Supportive vs. Control	-0.026	-0.028	-0.025	-0.026
	(0.018)	(0.017)	(0.017)	(0.017)
Opposing vs. Control	-0.054***	-0.053***	-0.050***	-0.051***
	(0.018)	(0.017)	(0.017)	(0.017)
Opposing vs. Supportive	-0.027	-0.024	-0.025	-0.024
	(0.018)	(0.017)	(0.017)	(0.017)
Observations	3,868	3,868	3,868	3,868
R^2	0.0023	0.1001	0.1049	0.1069
C. Newsletter Sign-Up				
Supportive vs. Control	0.015	0.012	0.012	0.012
	(0.016)	(0.015)	(0.015)	(0.015)
Opposing vs. Control	0.002	0.007	0.008	0.008
	(0.016)	(0.015)	(0.015)	(0.015)
Opposing vs. Supportive	-0.013	-0.006	-0.004	-0.004
	(0.016)	(0.015)	(0.015)	(0.015)
Observations	3,868	3,868	3,868	3,868
R^2	0.0003	0.1126	0.1187	0.1199

Notes: This table reports estimates from equation (2) with separate indicators for the Supportive and Opposing conditions. The omitted category is the No Comments control. Column (1) includes no covariates. Column (2) is the main specification, and includes the pre-specified baseline controls: age, gender, education, race, ethnicity, party affiliation, political views, baseline racial attitudes, and beliefs about others' views. Column (3) selects covariates from the full candidate set by LASSO. Column (4) includes the full candidate set without selection. Panel A reports the post-exposure attitudes index, aggregated by inverse-covariance weighting following Anderson (2008) and standardized to mean zero and standard deviation one in the control group, with higher values indicating greater alignment with the organization's position. Panel B reports *Donated*, an indicator for committing a positive amount in the incentivized donation task, and Panel C reports *Newsletter Sign-Up*, an indicator for providing an email address to sign up for a newsletter. Standard errors robust to heteroskedasticity are in parentheses. $^{*}p < 0.10$; $^{**}p < 0.05$; $^{***}p < 0.01$.

F.3 Results on Pre-specified Secondary Outcomes

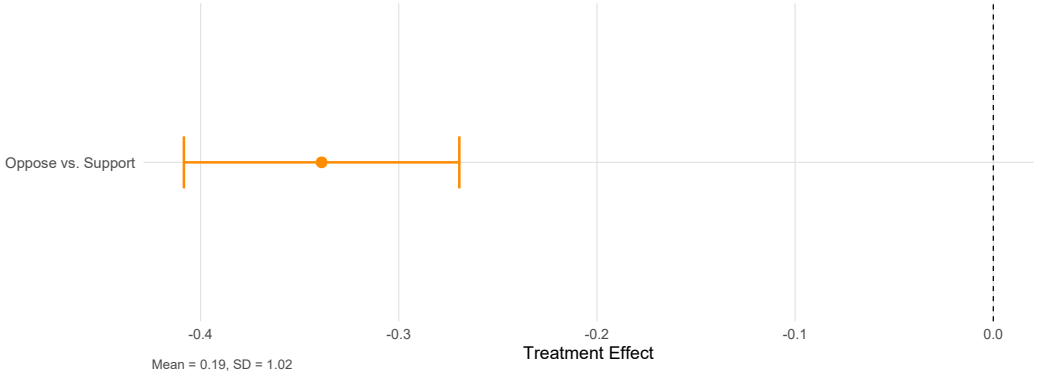
This appendix reports results on pre-specified secondary outcomes. Figure F13 reports treatment effects on secondary attitudinal outcomes. The pattern mirrors the primary attitude index: opposing comments produce negative but imprecisely estimated effects on the index, driven primarily by reduced favorability towards Black Lives Matter. Effects on perceived importance of voting and technology issues — which are not directly covered by the stimuli — are small and close to zero. Supportive comments have little effect across all components. Figure F14 shows that the secondary curiosity index follows a similar pattern to the primary curiosity and reflection index, with opposing comments reducing curiosity relative to supportive comments. Turning to beliefs about commenters, Figure F15 shows that supportive comments lead respondents to name progressives as the group more likely to comment on the post, while opposing comments have the opposite effect, consistent with respondents reading the stance of the comments they saw. Finally, Figure F16 shows that exposure to a comment section, regardless of stance, leads respondents to name men as the group more likely to comment, relative to the control, even though each post displayed one comment under a male name and one under a female name.

Figure F13: Treatment Effects on Secondary Attitude Index



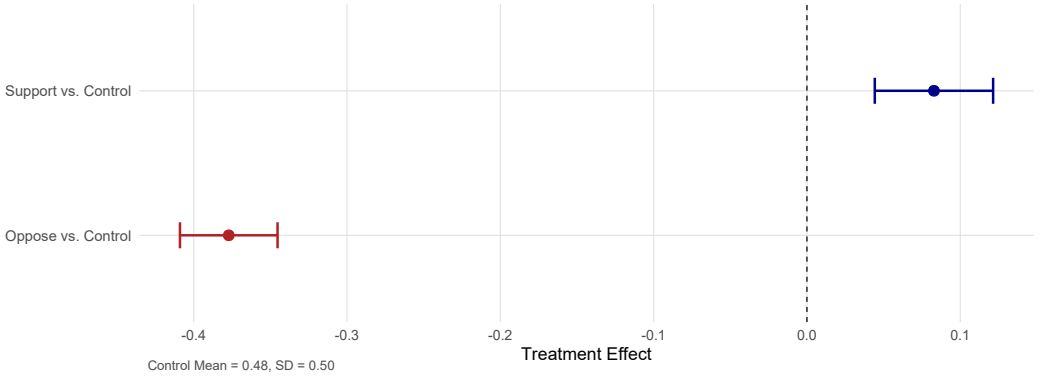
Notes: This figure decomposes treatment effects on post-exposure attitudes (standardized, control mean = 0.00, SD = 1.00) into its component outcomes: favorability toward Black Lives Matter, and perceived importance of two racial justice sub-issues — voting and technology — that are not directly covered by the intervention posts. Effects are shown separately for the Oppose vs. Control and Support vs. Control comparisons, with the index estimate shown in bold and individual outcomes indented below. Horizontal lines represent 95% confidence intervals.

Figure F14: Treatment Effects on Secondary Curiosity Index



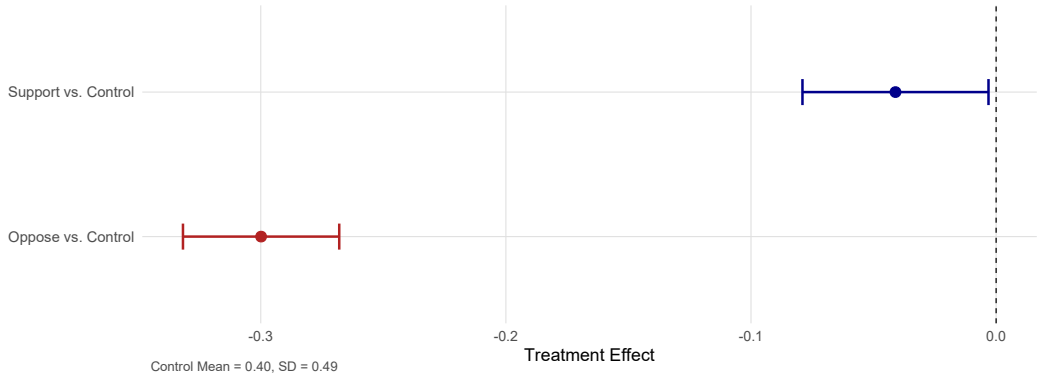
Notes: This figure reports treatment effects on a secondary curiosity index combining two measures: the extent to which comments increase curiosity about the underlying topic, and the extent to which they increase curiosity about the organization. The effect is shown for the Oppose vs. Support comparison. The mean and standard deviation shown below the figure are for the Supportive arm. Horizontal lines represent 95% confidence intervals.

Figure F15: Treatment Effects on Perceived Progressiveness of Commenters



Notes: This figure reports treatment effects on respondents' beliefs about the political ideology of commenters. Respondents were asked which group they believe is more likely to comment on the post; the outcome is coded as one if the respondent perceives progressives as more likely to comment. Effects are shown for the Oppose vs. Control and Support vs. Control comparisons. Horizontal lines represent 95% confidence intervals.

Figure F16: Treatment Effects on Perceived Gender of Commenters



Notes: This figure reports treatment effects on respondents' beliefs about the gender of commenters. Respondents were asked which group they believe is more likely to comment on the post; the outcome is coded as one if the respondent perceives women as more likely to comment. Effects are shown for the Oppose vs. Control and Support vs. Control comparisons. Horizontal lines represent 95% confidence intervals.

G Cost-Benefit Analysis for Fundraising Campaigns

We consider a nonprofit that chooses whether to tolerate opposing comments below its ads. We focus on opposing comments because they deliver the sharpest organizational trade-off in our setting: they increase clicks and website traffic, but reduce donations and shift attitudes in a less progressive direction. Let $a \in \{C, O\}$ denote the comment policy, where C is the control policy (no comments), and O is the opposing-comments policy. The organization is assumed to maximize expected donations generated by a campaign with budget B:

$$\mathcal{D}_a = B \times r_a \times CTR_a \times CVR_a, \tag{7}$$

where r_a is the number of users reached per dollar spent, CTR_a is the click-through rate out of reached users, and CVR_a is the donation conversion rate conditional on click.

Taking the ratio of donations under opposing comments relative to the control gives

$$\frac{\mathcal{D}_O}{\mathcal{D}_C} = \underbrace{\frac{r_O}{r_C}}_{\kappa} \times \underbrace{\frac{CTR_O}{CTR_C}}_{\tau} \times \underbrace{\frac{CVR_O}{CVR_C}}_{\text{conversion effect}} \tag{8}$$

We decompose the conversion effect into two components. First, we use the survey experiment to proxy for the direct downstream effect of opposing comments on donation decisions once users have already processed the ad. Let

$$\delta \equiv \frac{DonationRate_O}{DonationRate_C} \tag{9}$$

Using the estimates in our survey experiment,

$$\delta = \frac{0.677}{0.73} = 0.927 \tag{10}$$

Second, we introduce a reduced-form "traffic quality" parameter, q, which captures any additional change in conversion arising from the composition of users who click or are reached. In particular, $q < 1$ if opposing comments attract lower-intent users, or if the platform's delivery algorithm shifts exposure toward users who are more likely to engage but less likely to donate. Conversely, $q > 1$ if opposing comments attract or reach users who are more likely to donate.

Combining these pieces,

$$\frac{CVR_O}{CVR_C} = \delta \times q, \tag{11}$$

so that

$$\frac{\mathcal{D}_O}{\mathcal{D}_C} = \kappa \times \tau \times \delta \times q. \tag{12}$$

The parameter κ captures the extent to which opposing comments reduce advertising costs and therefore increase reach per dollar. Under our benchmark specification, we set $\kappa = 1$, which corresponds to the case in which delivery efficiency is unaffected. This is a natural starting point given

our experimental design, which imposed fixed budgets and aimed to keep reach balanced across arms. Once these constraints are relaxed, however, delivery efficiency may change. If the platform treats engagement as a positive signal and reduces cost per reach under opposing comments, then $\kappa > 1$. If instead opposing comments make delivery less efficient, then $\kappa < 1$. In our scenario analysis, we vary κ only in a narrow range around one in order to remain conservative and to reflect the fact that large delivery-cost differences are not part of our benchmark.

The parameter τ is directly computed from the Facebook field experiment:

$$\tau = \frac{CTR_O}{CTR_C} = \frac{0.261\%}{0.228\%} = 1.145. \tag{13}$$

Substituting (10) and (13) into (12) yields

$$\frac{\mathcal{D}_O}{\mathcal{D}_C} = 1.145 \times 0.927 \times \kappa \times q = 1.061 \times \kappa \times q. \tag{14}$$

Equation (14) is our main back-of-the-envelope formula. It shows that the effect of tolerating opposing comments depends on two parameters:

- κ: a narrow cost-efficiency parameter, centered at one, capturing small deviations from the fixed-budget/fixed-reach benchmark;
- q: a traffic-quality parameter, capturing whether the users induced to click or reached by the platform are more or less likely to convert.

Base case In the benchmark case, we set $\kappa = 1$, $q = 1$, so that opposing comments affect donations only through the observed click effect in Facebook and the direct downstream donation effect in the survey experiment. In this case,

$$\frac{\mathcal{D}_O}{\mathcal{D}_C} = 1.061, \tag{15}$$

implying a 6.1% increase in expected donations.

Scenario analysis To assess sensitivity to deviations in the other parameters, we vary κ only slightly around one and allow for larger, asymmetric movements in q. Specifically, we center the analysis at $q = 1$, allow a modest upside case with $q > 1$, and place greater weight on $q < 1$ because our evidence suggests that opposing comments attract relatively more clicks from users in less progressive areas, making a deterioration in traffic quality more plausible than an improvement. In particular, we consider combinations of: $\kappa \in \{0.98, 1.00, 1.02\}$ and $q \in \{0.90, 1.00, 1.05\}$. Table G1 shows the changes in campaign donation rates under different scenarios.

The break-even condition is

$$\kappa \times q > \frac{1}{1.061} \approx 0.943. \tag{16}$$

Table G1: Changes in Campaign Donation Rates Under Alternative Scenarios

	Traffic-quality parameter q					
Cost-efficiency parameter κ	0.90		1.00		1.05	
0.98	0.936	(−6.4%)	1.040	(+4.0%)	1.092	(+9.2%)
1.00	0.955	(−4.5%)	1.061	(+6.1%)	1.114	(+11.4%)
1.02	0.974	(−2.6%)	1.082	(+8.2%)	1.136	(+13.6%)

Notes: Each cell reports the implied ratio $\mathcal{D}_O/\mathcal{D}_C = 1.061 \times \kappa \times q$. The benchmark case is $(\kappa, q) = (1, 1)$. We keep κ in a tight range around one because the experimental design split budgets evenly across arms and optimized for reach, so large delivery-cost differences are not part of the maintained benchmark. By contrast, q is allowed to vary more widely because opposing comments may change the quality of induced traffic and, if delivery responds to engagement, the composition of users reached. Values greater than one imply that tolerating opposing comments increases expected donations; values below one imply that it decreases expected donations.

Thus, under the benchmark $\kappa = 1$, a deterioration in traffic quality of only about 5.7% is enough to overturn the baseline gain from higher click-through rates.

Discussion The benchmark case $(\kappa, q) = (1, 1)$ suggests that the overall donation rate increases by 6.1%. This benchmark is useful as a reference point, but it likely corresponds to a relatively optimistic scenario in which the additional traffic generated by opposing comments has the same propensity to donate as baseline traffic. Our results suggest that this assumption may be unrealistic: opposing comments generate relatively more traffic from users in less progressive areas, making it plausible that the induced traffic is of lower quality for fundraising purposes. For this reason, cases with $q < 1$ are likely to be more informative than the benchmark case, even though we allow both parameters to vary in the scenario analysis.

This exercise also abstracts from other potential objectives of the organization. In particular, we do not incorporate effects on attitudes, newsletter sign-ups, or shifts in the platform's delivery algorithm, even though these may also matter for advocacy organizations. We focus on donations because they map naturally into a dollar-valued objective and therefore allow for a simple back-of-the-envelope comparison. The table should therefore be interpreted as a stylized exercise focused on fundraising. Nonetheless, it highlights that modest gains in traffic can be offset—or overturned—if opposing comments attract lower-intent users or shift delivery toward users who are less likely to convert.

H Survey Instrument

Screening

Welcome! We have a few quick questions before we start.

This should take no more than 20 seconds. We will let you know if you are eligible for the study and provide details on participation payments.

1. Do you live in the United States?
   [Yes / No]

2. What is your age? *[number entry box]*

3. What is your gender?
   [Male / Female / Non-binary / other / I prefer not to answer]

4. What is your race?
   [American Indian / Alaska Native / Asian / Pacific Islander / Black / African American / White / Other / Mixed Race]

5. Are you of Hispanic or Latino origin?
   [Yes / No]

[Continue if U.S. resident, aged 18–64.]

Consent

[Consent form]

I agree to participate, and I promise to read the questions carefully and answer honestly

I do not agree to participate, or I cannot promise to read the questions carefully and answer honestly

Baseline Opinions

Thanks for agreeing to participate! We value your opinions and are interested in hearing what you think.

1. In politics, as of today, do you consider yourself a Republican, a Democrat or an independent?
   [Republican / Democrat / Independent / Other]

2. We hear a lot of talk these days about liberals and conservatives. Which of the following best describe your political view?
   [Very liberal / Liberal / Moderate; middle of the road / Conservative / Very conservative / Haven't thought much about this/don't know]

3. When it comes to giving African Americans equal rights with white Americans, do you think our country has...
   [Gone too far / Not gone far enough / Been about right]

   [Randomize order of "Gone too far" and "Not gone far enough."]

4. Do you believe that the increased public attention to the history of slavery and racism is generally good or bad for our society?
   [Very good / Somewhat good / Neither good nor bad / Somewhat bad / Very bad]

5. **On the issues of race and racism,** my position is...
   [Very progressive / Progressive / Moderate; middle of the road / Conservative / Very conservative / Haven't thought much about this / don't know]

   [Randomly flip the order.]

6. Now we'd like to know your best guess about how people in a **representative sample of adults in the United States** answered this question in 2025.

   "When it comes to giving African Americans equal rights with white Americans, do you think our country has..."

   Please estimate what percentage of respondents chose each response. Your answers should add up to 100%.

   [Constant sum question; entries for: ___% Gone too far / ___% Not gone far enough / ___% Been about right; must sum to 100.]

Social Media Use

1. How much time do you spend on social media (e.g., Facebook, Instagram, TikTok, YouTube) excluding Messenger and WhatsApp, on an average day?
   [Less than 5 minutes a day / Between 5 and 30 minutes a day / Between 30 and 60 minutes a day / Between 1 and 2 hours / Between 2 and 4 hours / More than 4 hours]

2. How often do you read or check comments on social media?
   [Never / Rarely / Sometimes / Often / Very often]

Additional Demographics

1. In what zip code do you currently live? *[text entry box; validation: U.S. ZIP code]*

2. What is the highest degree or level of schooling that you have completed?
   [Less than a high school diploma / High school diploma or equivalent (for example: GED) / Some college but no degree / Associate's degree / Bachelor's degree / Graduate degree (for example: MA, MBA, JD, PhD)]

Attention Check

1. In order to facilitate our research, we are interested in knowing certain factors about you. Specifically, we are interested in whether you actually take the time to read the instructions; if not, then the data we collect based on your responses will be invalid. So, in order to demonstrate that you have read the instructions, please ignore the next question, and simply write "I read the instructions" in the "Any comments?" box below. Thank you very much.

   What is your marital status?
   [Single / Married / Other]

   Any comments? *[text box]*

Intervention

[Participants randomized into 3 groups: No Comments, Supportive, Opposing.]

[3 posts of the same treatment type are shown, varying order of topics: education, environment, and police.]

We are interested in your reactions to social media posts about racial justice. You will be shown three posts from Color of Change, a leading U.S.-based racial justice advocacy organization.

[Next page.]

Here is a post from Color of Change:

[Screenshot shown.]

[Next page.]

Now we will ask you some questions about this post.

[Post shown again.]

1. Which of the following emojis would you react with if you saw this post on Facebook?
   [Like / Love / Care / Haha / Wow / Sad / Angry / I would not react]

2. Would you comment on this post if you saw it on Facebook?
   [Yes / No]

3. *[If yes to previous question]* What comment would you make? *[open text]*

4. Would you click on this post to visit the website if you saw it on Facebook?
   [Yes / No]

[Screenshots of remaining posts shown; questions above repeated for each post.]

Newsletter

1. Would you like to sign up for the Color of Change newsletter?
   [Yes / No]

Donation

1. You have been automatically enrolled in a lottery to win up to $100. If you win, you have the option to donate some or all of your winnings to Color of Change.

   The payment will be made to you as a bonus, so no further action is required on your part. If you are one of the lottery winners, you will be paid, in addition to your participation payment, $100 minus the amount you donated. We will directly pay your desired donation amount to Color of Change.

   How much, if any, would you be willing to donate to Color of Change in case you won $100?
   [number entry]

Post-Exposure Opinion

1. How would you describe your overall opinion of Color of Change?
   [Very unfavorable / Somewhat unfavorable / Neutral / Somewhat favorable / Very favorable]

2. How would you describe your overall opinion of Black Lives Matter?
   [Very unfavorable / Somewhat unfavorable / Neutral / Somewhat favorable / Very favorable]

3. How willing would you be to discuss political issues with someone who has **progressive views** on racial issues?
   [Very unwilling / Somewhat unwilling / Neither willing nor unwilling / Somewhat willing / Very willing]

4. How willing would you be to discuss political issues with someone who has **conservative views** on racial issues?
   [Very unwilling / Somewhat unwilling / Neither willing nor unwilling / Somewhat willing / Very willing]

5. People differ in how important they consider different racial justice issues.

   How important is each of the following issues to you personally?
   [Matrix; rows: Voter suppression and voting rights / Criminal-justice reform (e.g., policing, sentencing, incarceration) / Education equity (e.g., school funding, achievement gaps) / Environmental justice (e.g., pollution exposure, clean air/water access) / Technology fairness (e.g., algorithmic bias, digital discrimination); 5-point scale: Not at all important / Slightly important / Moderately important / Very important / Extremely important]

6. Please tell us the extent to which you agree or disagree with the statement below.

   It's really a matter of some people not trying hard enough. Black people could be just as well off as white people if they would only try harder.
   [Strongly disagree / Disagree / Slightly disagree / Neither agree nor disagree / Slightly agree / Agree / Strongly agree]

7. What do you think is the most important issue facing Black people today? *[open text]*

Opinion about Others

Now we are going to ask you to make a guess.

You can earn a Guess Bonus of up to $1 based on the accuracy of your estimate. We will compare your estimate to the actual percentages observed in this study. The formula we use rewards you more when your estimate is closer to the true value. **The closer your guess is to the correct percentage, the larger your bonus.**

Your best strategy is to give your honest, best estimate.

(You do *not* need to know the formula to earn the bonus.)

[Next page.]

1. What percentage of participants in this U.S. adult sample do you think **agreed or strongly agreed** with the following statement?

   "It's really a matter of some people not trying hard enough. Black people could be just as well off as white people if they would only try harder."

   Please enter a number between **0 and 100**.

   You can earn a **bonus** based on how close your estimate is to the true value.

   ___% *[number entry]*

   [Pop-up window with quadratic scoring rule: Guess Bonus = $1 − ((Your Answer − True Value)/100)^2. Small errors reduce your bonus slightly; larger errors reduce it much more. If your estimate is exactly correct, you will earn $1. Your best strategy is to give your honest, best estimate.]

[Environment post shown again.]

2. How thought-provoking do you find this post?
   [Not at all / A little / Somewhat / Very / Extremely]

Opinion about Comments in Post [treatment arms only]

Now consider the **comments** below the following social media post.

[Environment post with comments shown again.]

1. To what extent do the comments make you feel:
   [Matrix; rows: Angry / Annoyed; 5-point scale: Not at all / A little / Somewhat / Very / Extremely]

2. How thought-provoking do you find these comments?
   [Not at all / A little / Somewhat / Very / Extremely]

3. To what extent do the comments make you feel:
   [Matrix; rows: Curious about the topic / Curious about the organization; 5-point scale: Not at all / A little / Somewhat / Very / Extremely]

4. How representative do you think these comments are of what people generally think about this issue?
   [Not at all representative / Slightly representative / Moderately representative / Very representative / Extremely representative]

Opinion about Comments in Post [all arms]
[Environment post shown again.]

1. How interested would you be in opening the comment section to see more comments on this post?
   [Not at all interested / Slightly interested / Moderately interested / Very interested / Extremely interested]

2. In your opinion, which group is more likely to comment on this post?
   [Men / Women / Men and women are equally likely / Not sure]

3. In your opinion, which group is more likely to comment on this post?
   [Progressives / Conservatives / Both groups are equally likely / Not sure]

Opinion about Comments in General

1. What are the main reasons you read or look at comments on social media? *[open text]*

Newsletter Sign-Up
[Shown only to participants who answered Yes to the newsletter question.]
You said that you like to sign up for the Color of Change newsletter.

1. Please enter your email address below. Your email will only be shared with Color of Change, and only for the purpose of subscribing you to their newsletter.
   [text entry box]

AI Use

1. Did you use AI at all to help you fill out this survey?
   [Yes / No]

2. *[If yes]* What question did you use AI to help you answer? *[open text]*

Final Feedback

Thank you! We really appreciate you for participating in this research!

Please let us know if you have any other feedback.

Make sure to click the next arrow to submit your survey responses.

[text box]

I Codebook for NLP-Based Measures

This appendix documents the NLP-based measures used in the paper. For all GPT-based classifications, we use GPT-4 with temperature set to 0. Toxicity is measured separately using the Google Perspective API.

I.1 GPT-Based Conversation Measures

Reciprocity, Justification, and Respect. These three measures are coded once for each direct comment. The input consists of the original Facebook post, one direct comment, and all replies to that comment, so that the comment is assessed in the context of the exchange it started. GPT-4 is instructed to return a JSON object with three binary indicators:

- **Reciprocity** equals 1 when participants engage the existing conversation and remain on topic.
- **Justification** equals 1 when participants who advocate a position provide reasons or arguments.
- **Respect** equals 1 when the exchange remains civil and free of threats, insults, humiliating language, or silencing expressions.

I.2 GPT-Based Comment-Level Measures

Political Stance (1–5). GPT-4 classifies each comment on a five-point ideological scale relative to the issue discussed in the ad:

1. Strongly progressive or left-leaning
2. Slightly or moderately progressive
3. Centrist, unclear, or no explicit stance
4. Slightly or moderately conservative
5. Strongly conservative or right-leaning

Political Stance (binary). A second GPT-4 measure collapses ideological stance into a binary variable:

- 1 = progressive or left-leaning
- 2 = conservative or right-leaning

When the comment is neutral or ambiguous, GPT-4 is instructed to choose the closest side based on the language used.

Sentiment. GPT-4 classifies each comment from 1 to 5 according to its sentiment toward the main statement of the ad:

1. Highly negative
2. Somewhat negative
3. Neutral or mixed
4. Somewhat positive
5. Very positive

Informativeness. GPT-4 classifies each comment from 1 to 3 according to how much useful or fact-based information it provides:

1. Not informative
2. Somewhat informative
3. Highly informative

Offensiveness. GPT-4 classifies each comment from 1 to 3 according to the presence of offensive or insulting language:

1. Not offensive
2. Mildly offensive
3. Highly offensive

Agreement. GPT-4 classifies each comment from 1 to 3 according to whether it agrees with the message of the ad:

1. Clear agreement
2. Neutral or unclear
3. Clear disagreement

I.3 API-Based Measure

Toxicity Score. Toxicity is measured using the Google Perspective API, requesting the `TOXICITY` attribute for each comment. The API returns a continuous score between 0 and 1, where higher values indicate a higher likelihood that the comment is perceived as toxic.